US005469354A

# United States Patent [19]

Hatakeyama et al.

[11] Patent Number: 5,469,354

[45] Date of Patent: Nov. 21, 1995

[54] **DOCUMENT DATA PROCESSING METHOD AND APPARATUS FOR DOCUMENT RETRIEVAL**

[75] Inventors: **Atsushi Hatakeyama**, Kokubunji; **Hiromichi Fujisawa**; **Kanji Kato**, both of Tokorozawa; **Hisamitsu Kawaguchi**, Sagamihara; **Naoki Minegishi**, Osaka; **Katsumi Tada**, Kokubunji; **Satoshi Asakawa**, Hirakata, all of Japan

[73] Assignee: **Hitachi, Ltd.**, Tokyo, Japan

[21] Appl. No.: **843,162**

[22] Filed: **Feb. 28, 1992**

**Related U.S. Application Data**

[63] Continuation-in-part of Ser. No. 555,483, Aug. 9, 1990, Pat. No. 5,168,533.

[30] **Foreign Application Priority Data**

Feb. 28, 1991 [JP] Japan ..... 3-058311
Dec. 25, 1991 [JP] Japan ..... 3-342695

[51] **Int. Cl.**[6] ..... **G06F 17/21**

[52] **U.S. Cl.** ..... **364/419.19**; 364/419.13; 364/419.07; 364/225.3

[58] **Field of Search** ..... 364/419.19, 419.13, 364/419.07, 225.3; 395/600

[56] **References Cited**

U.S. PATENT DOCUMENTS

| | | | |
|---|---|---|---|
| 4,870,568 | 9/1989 | Kahle et al. | 364/225.3 |
| 5,051,947 | 9/1991 | Messenger et al. | 364/956.1 |
| 5,168,533 | 12/1992 | Kato et al. | 364/200 |
| 5,206,949 | 4/1993 | Cochran et al. | 395/600 |
| 5,220,625 | 6/1993 | Hatakeyama et al. | 382/54 |

FOREIGN PATENT DOCUMENTS

| | | |
|---|---|---|
| 0437615A1 | 7/1991 | European Pat. Off. . |
| 63-198124 | 8/1988 | Japan . |
| 3-125263 | 5/1991 | Japan . |
| 5-55912 | 8/1993 | Japan . |
| 5-76068 | 10/1993 | Japan . |
| WO90/16036 | 2/1990 | WIPO . |

OTHER PUBLICATIONS

Mukhopadhyay et al., An Intelligent System for Document Retrievel in Distributed Office Environments, Journal of the American Society for Information Science, Jun. 17, 1985.

"State Machines Find the Pattern", System Design/Software, 8167 Computer Design, May 1985, No. 5, Littleton, Mass.

*Primary Examiner*—Robert A. Weinhardt
*Assistant Examiner*—Frantzy Poinvil
*Attorney, Agent, or Firm*—Antonelli, Terry, Stout & Kraus

[57] **ABSTRACT**

High-speed full document retrieval method and system capable of providing result of retrieval within practically acceptable short search time. Upon registration of documents in a document database, condensed texts are created by decomposing each of textual character strings of the documents to be registered into fragmental character strings in dependence on character species and by checking mutual inclusion relations existing among the fragmental character strings. A component character table is created in which characters occurring in each of the condensed texts are registered without duplication. The condensed texts and the component character table are registered in the data base together with the texts of the documents to be registered. Upon retrieval of a document containing a search term designated by a user, a component character table search is first executed to extract those documents which contain all species of characters constituting the search term by consulting the component character table, and subsequently a condensed text search is executed by consulting the condensed texts of the documents. Finally, a text body search is executed for extracting a document which satisfies query condition imposed on the search term by consulting the texts of the documents extracted through the component character table search and the condensed text search.

**50 Claims, 66 Drawing Sheets**

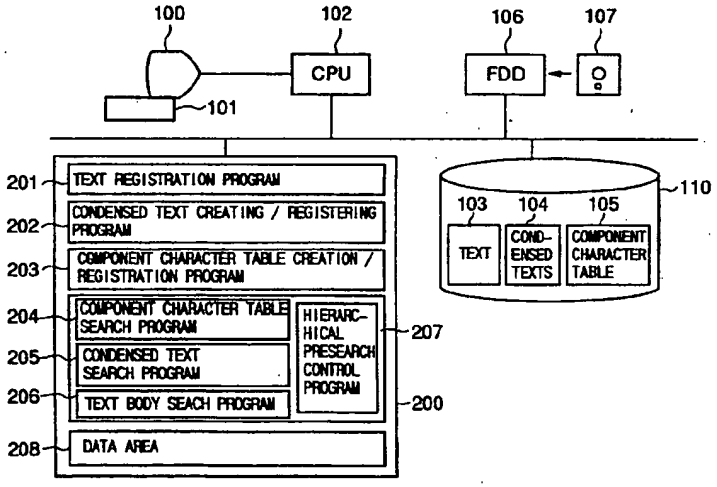


FIG. 1

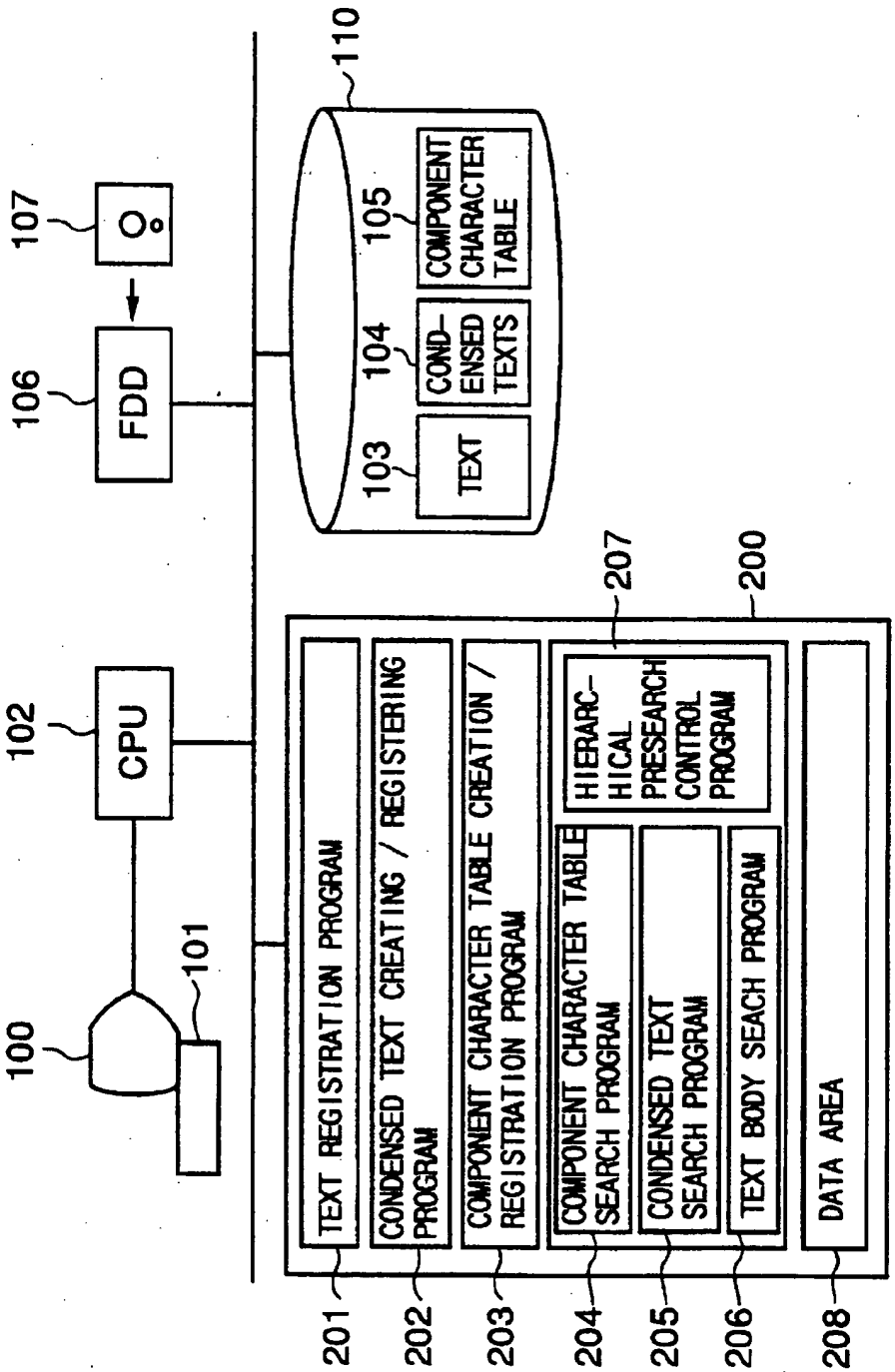

100
101
102
CPU
106
FDD
107
110
103
TEXT
104
COND-ENSED TEXTS
105
COMPONENT CHARACTER TABLE
200
207
TEXT REGISTRATION PROGRAM
CONDENSED TEXT CREATING / REGISTERING PROGRAM
COMPONENT CHARACTER TABLE CREATION / REGISTRATION PROGRAM
COMPONENT CHARACTER TABLE SEARCH PROGRAM
CONDENSED TEXT SEARCH PROGRAM
TEXT BODY SEACH PROGRAM
HIERARC-HICAL PRESEARCH CONTROL PROGRAM
DATA AREA
201
202
203
204
205
206
208

FIG. 2

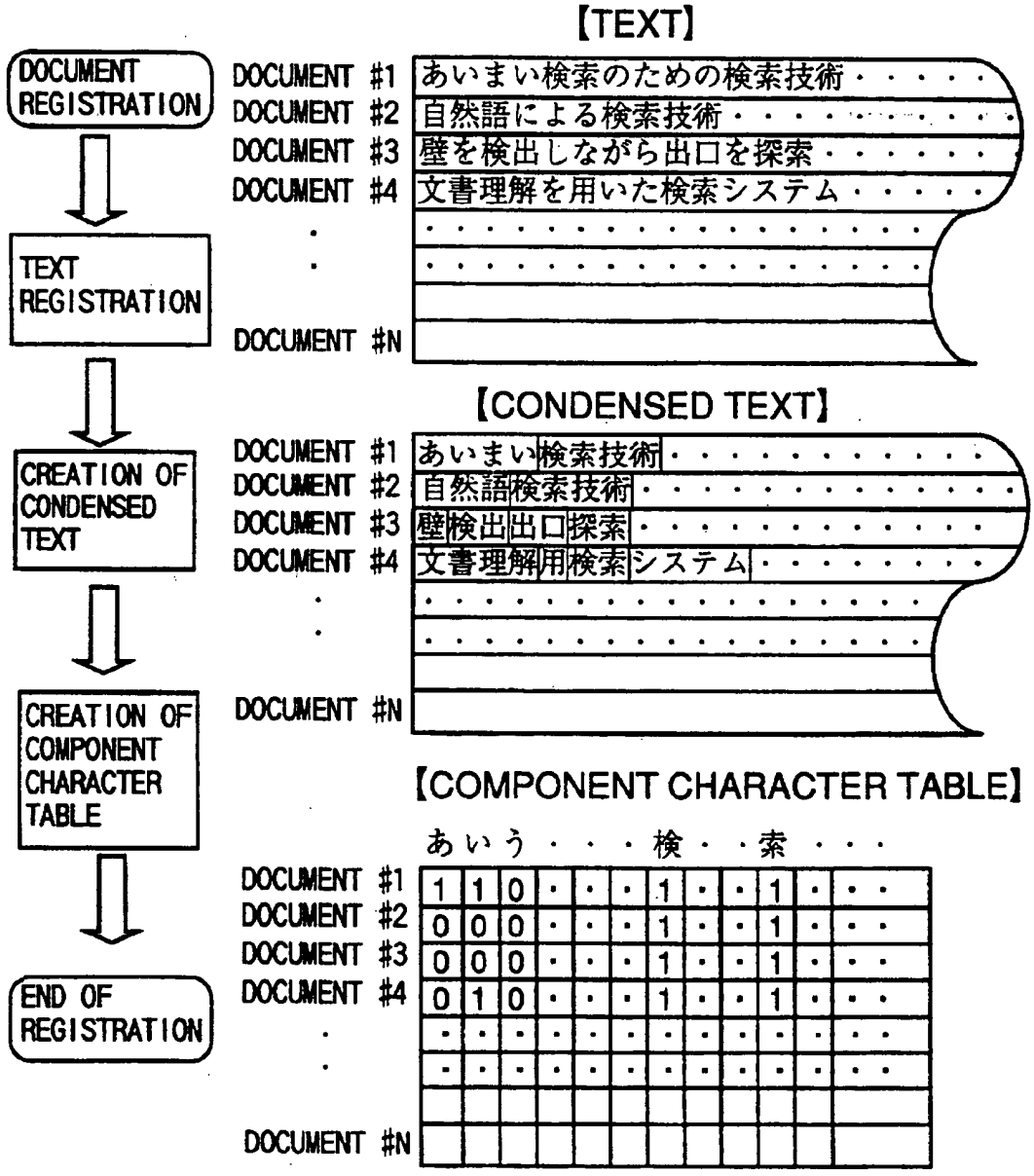

DOCUMENT REGISTRATION
TEXT REGISTRATION
CREATION OF CONDENSED TEXT
CREATION OF COMPONENT CHARACTER TABLE
END OF REGISTRATION
【TEXT】
DOCUMENT #1 あいまい検索のための検索技術
DOCUMENT #2 自然語による検索技術
DOCUMENT #3 壁を検出しながら出口を探索
DOCUMENT #4 文書理解を用いた検索システム
DOCUMENT #N
【CONDENSED TEXT】
DOCUMENT #1 あいまい 検索技術
DOCUMENT #2 自然語 検索技術
DOCUMENT #3 壁 検出 出口 探索
DOCUMENT #4 文書理解 用 検索 システム
DOCUMENT #N
【COMPONENT CHARACTER TABLE】
あ い う ・ ・ ・ 検 ・ ・ 索 ・ ・ ・
DOCUMENT #1 1 1 0 1 1
DOCUMENT #2 0 0 0 1 1
DOCUMENT #3 0 0 0 1 1
DOCUMENT #4 0 1 0 1 1
DOCUMENT #N

FIG. 3

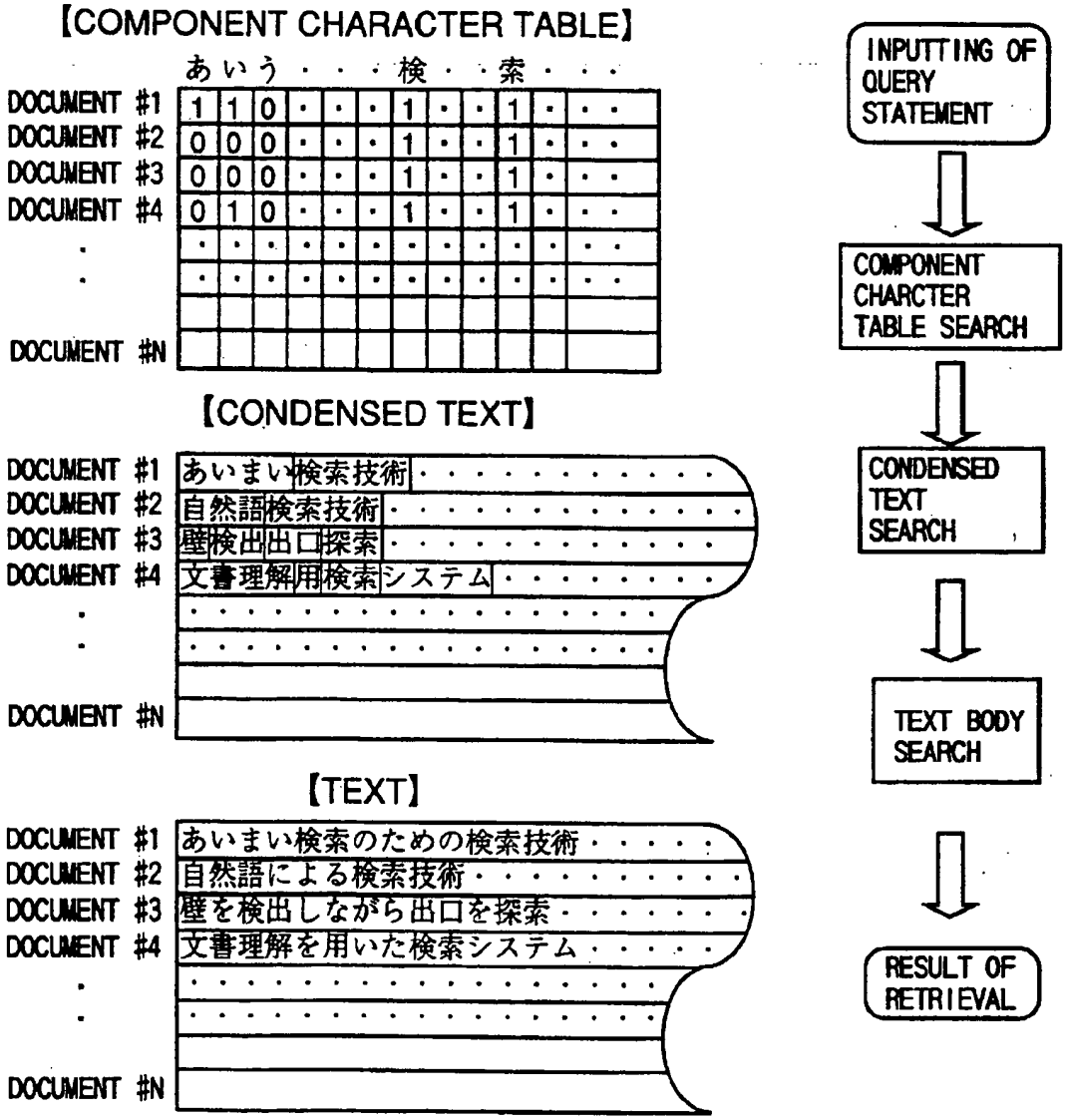

【COMPONENT CHARACTER TABLE】
あ い う ・ ・ ・ 検 ・ ・ 索 ・ ・ ・
DOCUMENT #1
DOCUMENT #2
DOCUMENT #3
DOCUMENT #4
DOCUMENT #N
【CONDENSED TEXT】
DOCUMENT #1 あいまい検索技術
DOCUMENT #2 自然語検索技術
DOCUMENT #3 壁検出出口探索
DOCUMENT #4 文書理解用検索システム
DOCUMENT #N
【TEXT】
DOCUMENT #1 あいまい検索のための検索技術
DOCUMENT #2 自然語による検索技術
DOCUMENT #3 壁を検出しながら出口を探索
DOCUMENT #4 文書理解を用いた検索システム
DOCUMENT #N
INPUTTING OF QUERY STATEMENT
COMPONENT CHARCTER TABLE SEARCH
CONDENSED TEXT SEARCH
TEXT BODY SEARCH
RESULT OF RETRIEVAL

FIG. 4

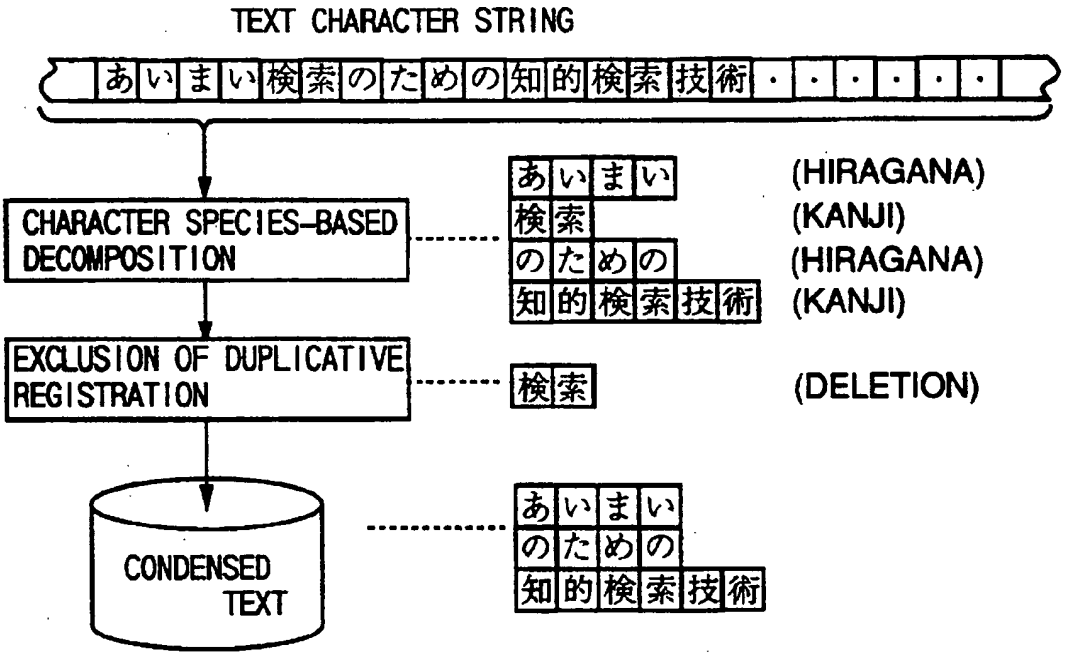

TEXT CHARACTER STRING
あいまい検索のための知的検索技術
CHARACTER SPECIES-BASED DECOMPOSITION
あいまい (HIRAGANA)
検索 (KANJI)
のための (HIRAGANA)
知的検索技術 (KANJI)
EXCLUSION OF DUPLICATIVE REGISTRATION
検索 (DELETION)
CONDENSED TEXT
あいまい
のための
知的検索技術

FIG. 5

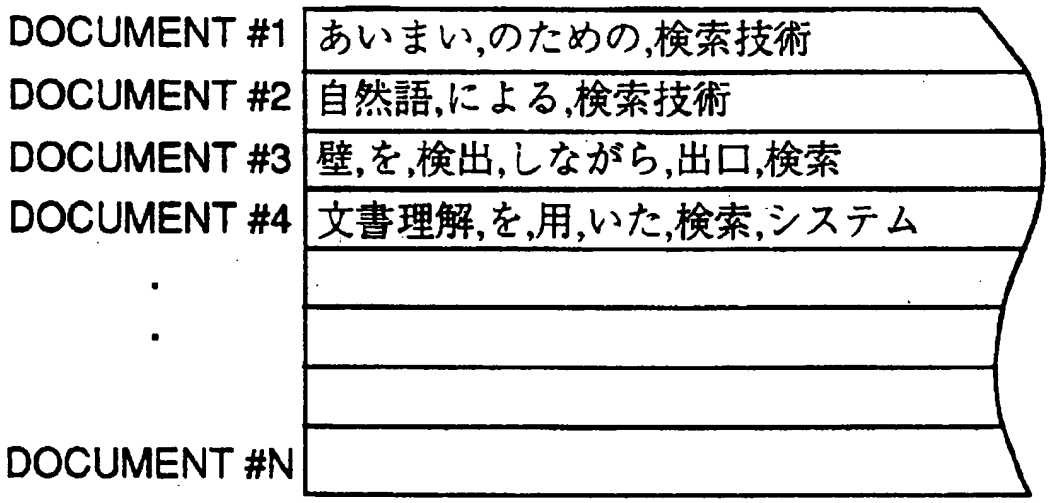

DOCUMENT #1
あいまい,のための,検索技術
DOCUMENT #2
自然語,による,検索技術
DOCUMENT #3
壁,を,検出,しながら,出口,検索
DOCUMENT #4
文書理解,を,用,いた,検索,システム
DOCUMENT #N

FIG. 6

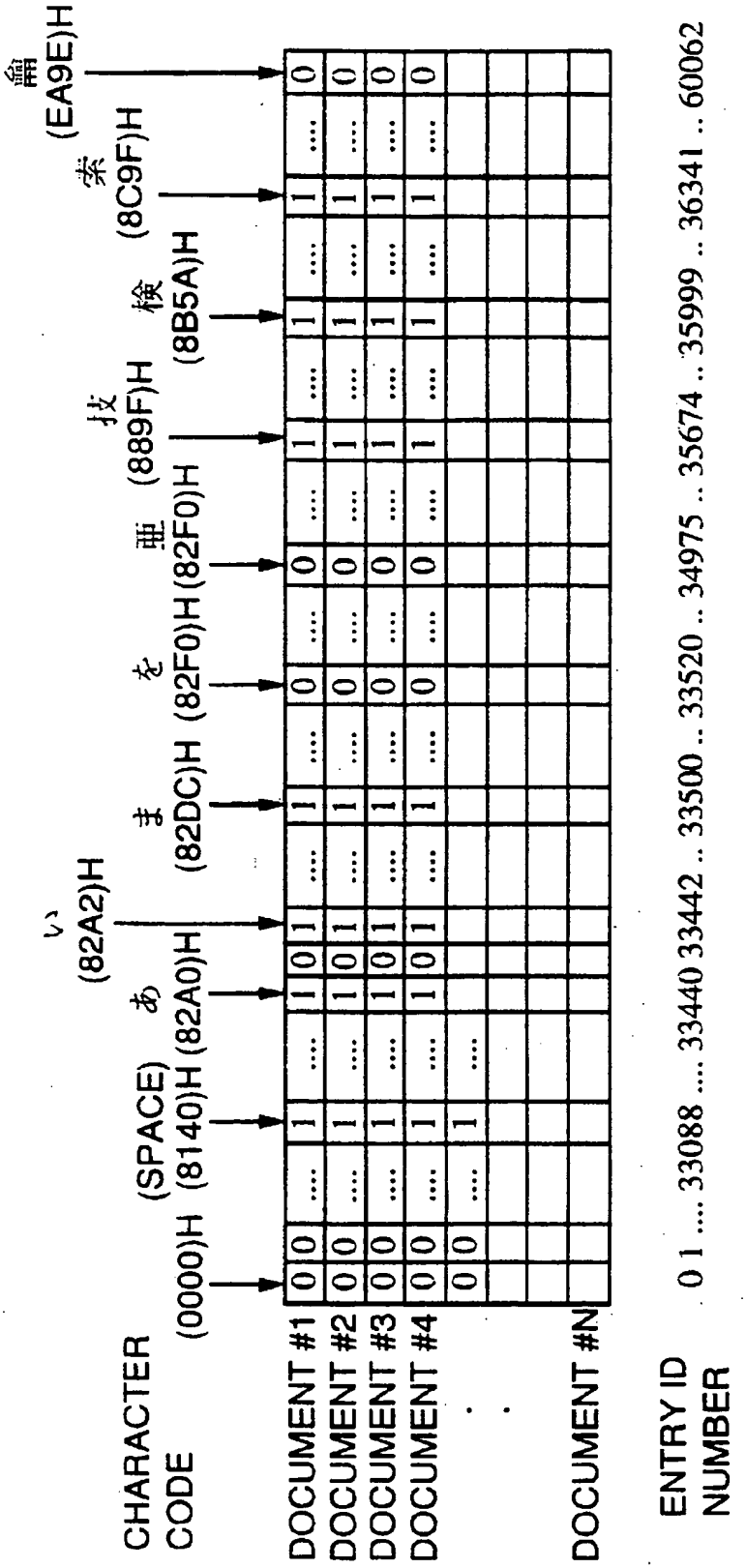

CHARACTER CODE
(0000)H
(SPACE) (8140)H
あ (82A0)H
い (82A2)H
ま (82DC)H
を (82F0)H
亜 (82F0)H
技 (889F)H
検 (8B5A)H
紫 (8C9F)H
論 (EA9E)H
DOCUMENT #1
DOCUMENT #2
DOCUMENT #3
DOCUMENT #4
DOCUMENT #N
ENTRY ID NUMBER
0 1 .... 33088 .... 33440 33442 .. 33500 .. 33520 .. 34975 .. 35674 .. 35999 .. 36341 .. 60062

FIG. 7

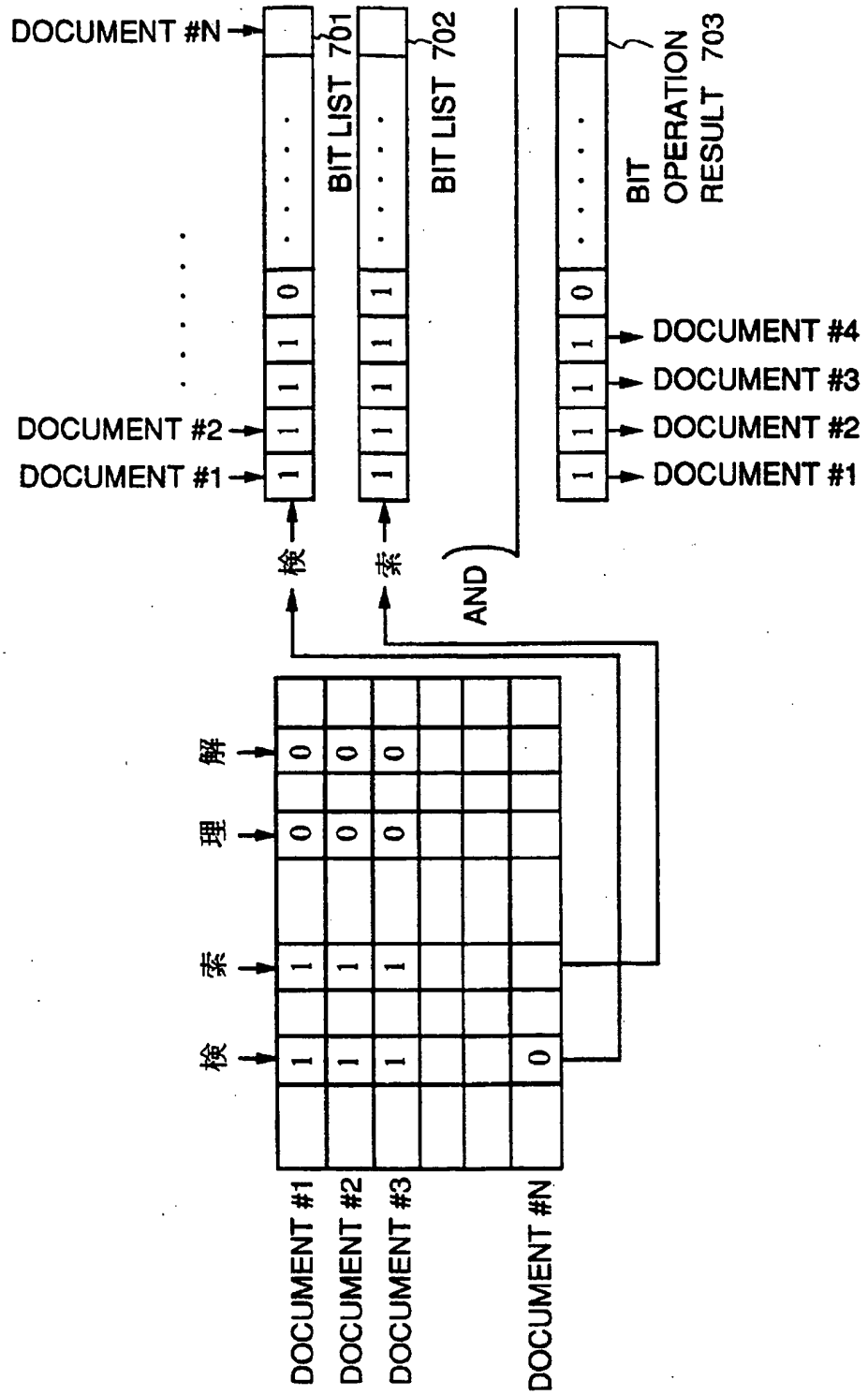

DOCUMENT #N
BIT LIST 701
BIT LIST 702
BIT OPERATION RESULT 703
DOCUMENT #4
DOCUMENT #3
DOCUMENT #2
DOCUMENT #1
DOCUMENT #2
DOCUMENT #1
検索
AND
解理
索
検
DOCUMENT #1
DOCUMENT #2
DOCUMENT #3
DOCUMENT #N

FIG. 8

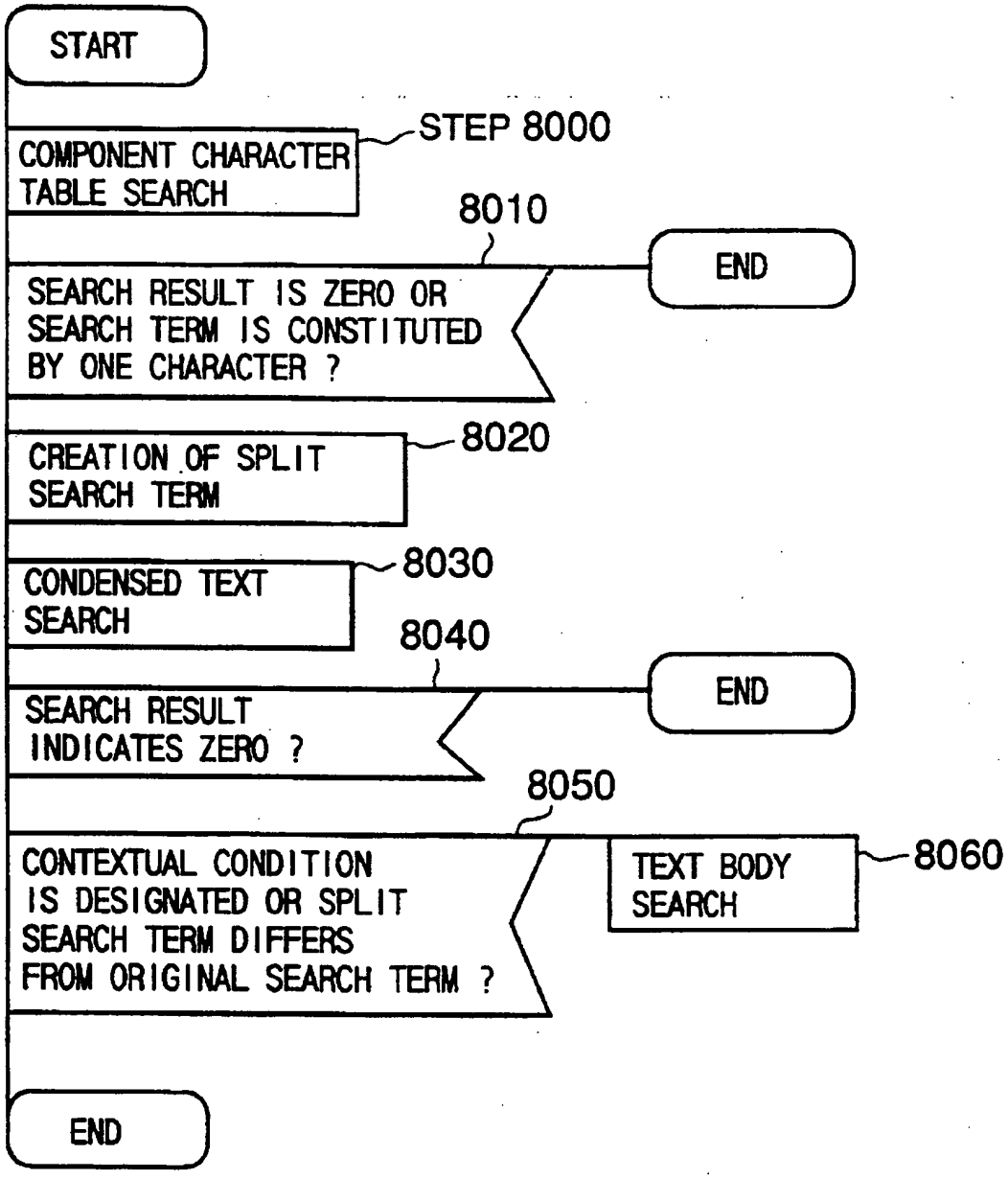

START
COMPONENT CHARACTER TABLE SEARCH
STEP 8000
8010
SEARCH RESULT IS ZERO OR SEARCH TERM IS CONSTITUTED BY ONE CHARACTER ?
END
CREATION OF SPLIT SEARCH TERM
8020
CONDENSED TEXT SEARCH
8030
8040
SEARCH RESULT INDICATES ZERO ?
END
8050
CONTEXTUAL CONDITION IS DESIGNATED OR SPLIT SEARCH TERM DIFFERS FROM ORIGINAL SEARCH TERM ?
TEXT BODY SEARCH
8060
END

FIG. 9

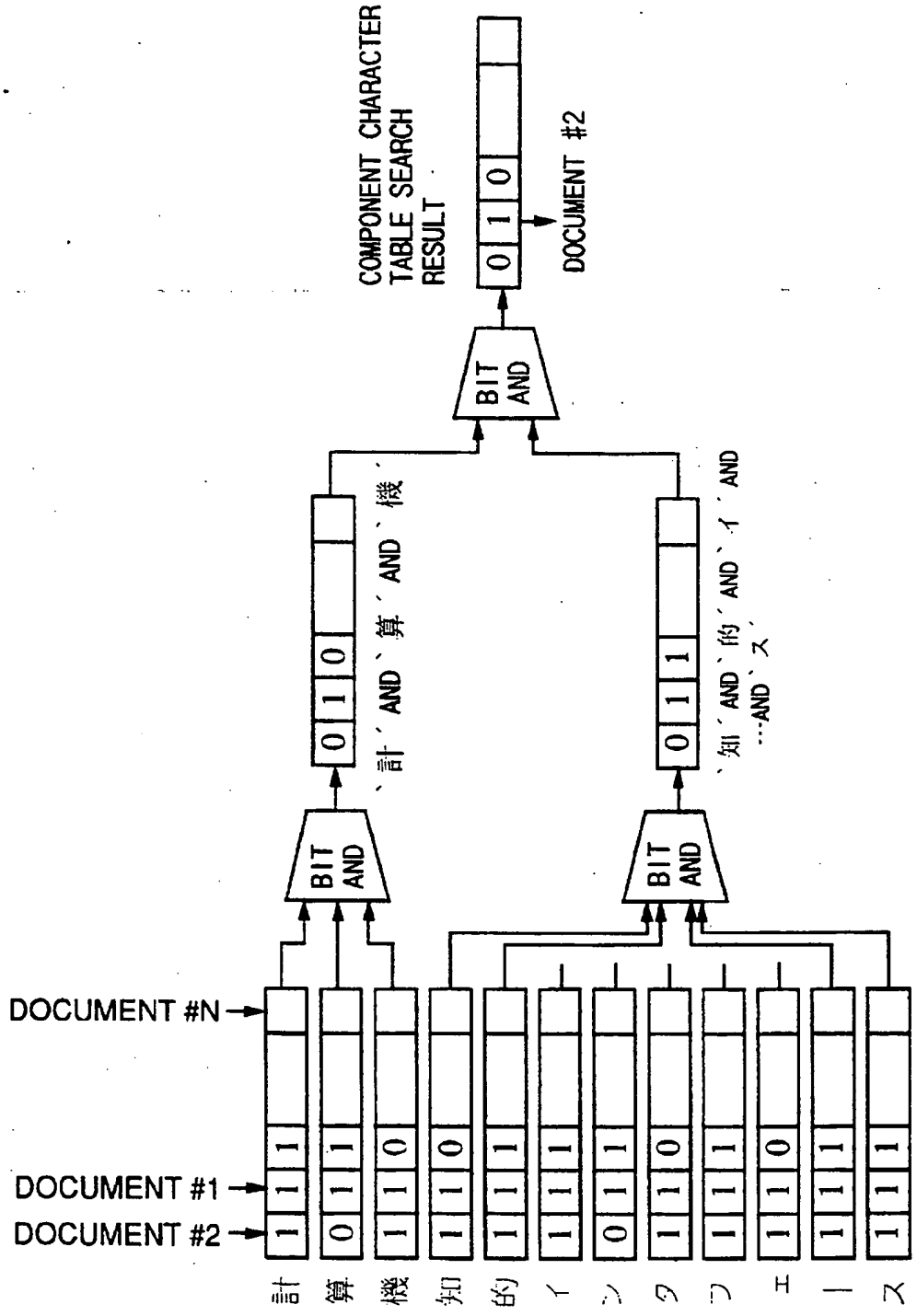

COMPONENT CHARACTER TABLE SEARCH RESULT
DOCUMENT #2
BIT AND
BIT AND
BIT AND
DOCUMENT #N
DOCUMENT #1
DOCUMENT #2
計
算
機
知
的
イ
ン
タ
フ
エ
ー
ス

FIG. 10

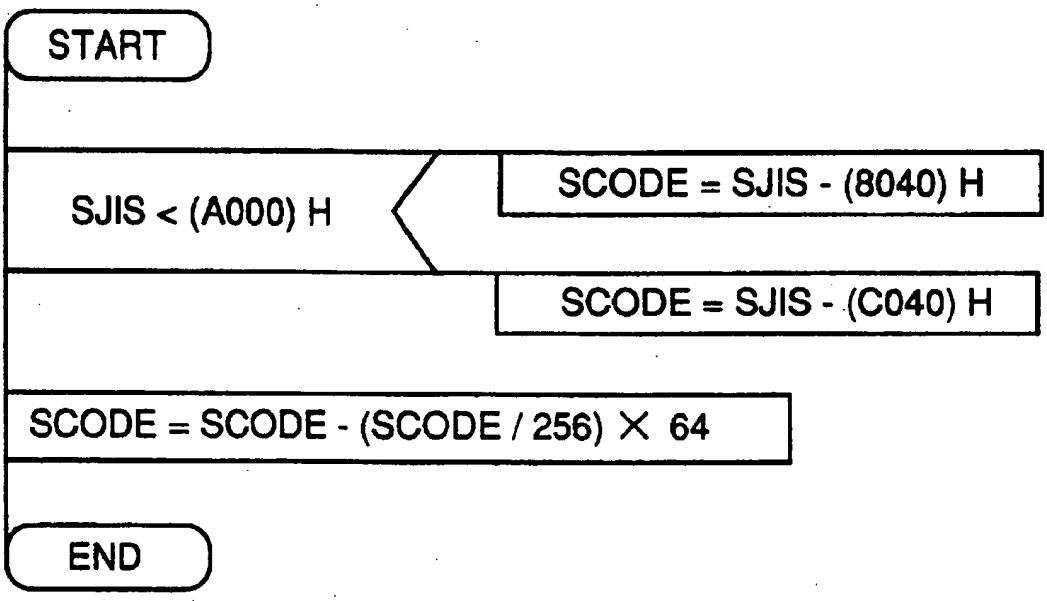

START
SJIS < (A000) H
SCODE = SJIS - (8040) H
SCODE = SJIS - (C040) H
SCODE = SCODE - (SCODE / 256) × 64
END

FIG. 11

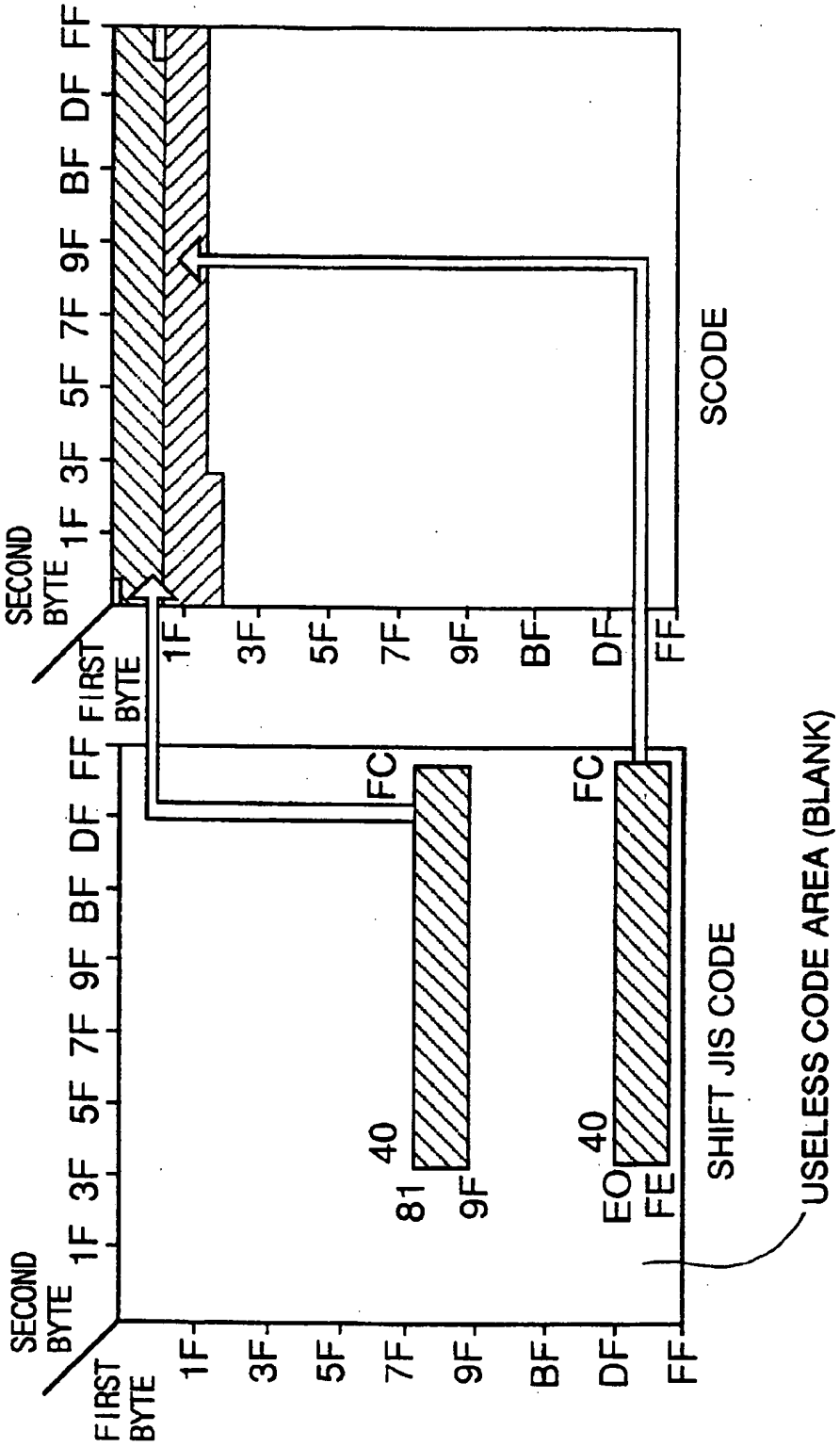

SECOND BYTE
FIRST BYTE
1F
3F
5F
7F
9F
BF
DF
FF
SCODE
FC
40
81
9F
EO
FE
SHIFT JIS CODE
USELESS CODE AREA (BLANK)

FIG. 12

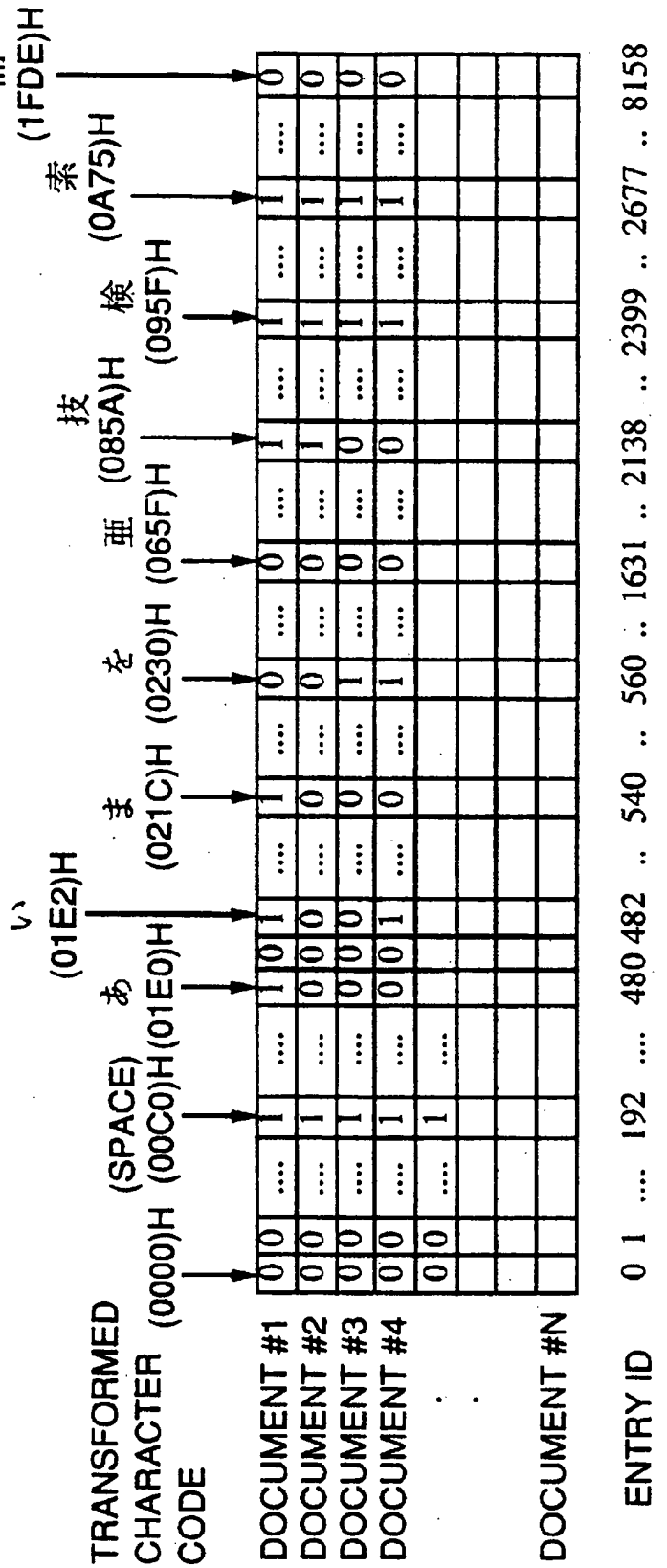

TRANSFORMED CHARACTER CODE
(0000)H
(SPACE) (00C0)H
あ (01E0)H
い (01E2)H
ま (021C)H
を (0230)H
亜 (065F)H
技 (085A)H
検 (095F)H
素 (0A75)H
傭 (1FDE)H
DOCUMENT #1
DOCUMENT #2
DOCUMENT #3
DOCUMENT #4
DOCUMENT #N
ENTRY ID NUMBER
0 1 .... 192 .... 480 482 .. 540 .. 560 .. 1631 .. 2138 .. 2399 .. 2677 .. 8158

FIG. 13

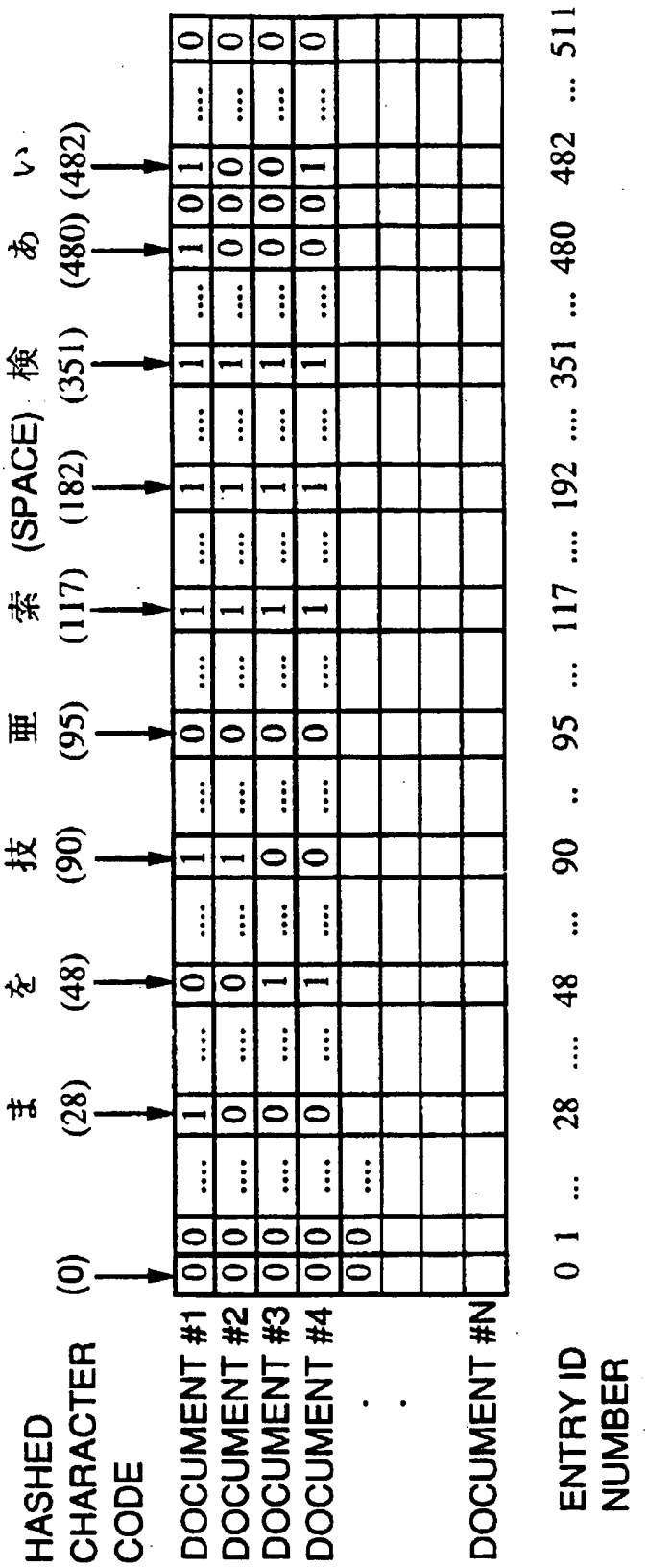

HASHED CHARACTER CODE
(0)
(28)
(48)
(90)
(95)
(117)
(182)
(SPACE)
(351)
(480)
(482)
DOCUMENT #1
DOCUMENT #2
DOCUMENT #3
DOCUMENT #4
DOCUMENT #N
ENTRY ID NUMBER
0 1 ... 28 .... 48 ... 90 .. 95 ... 117 .... 192 .... 351 ... 480 482 ... 511

FIG. 14

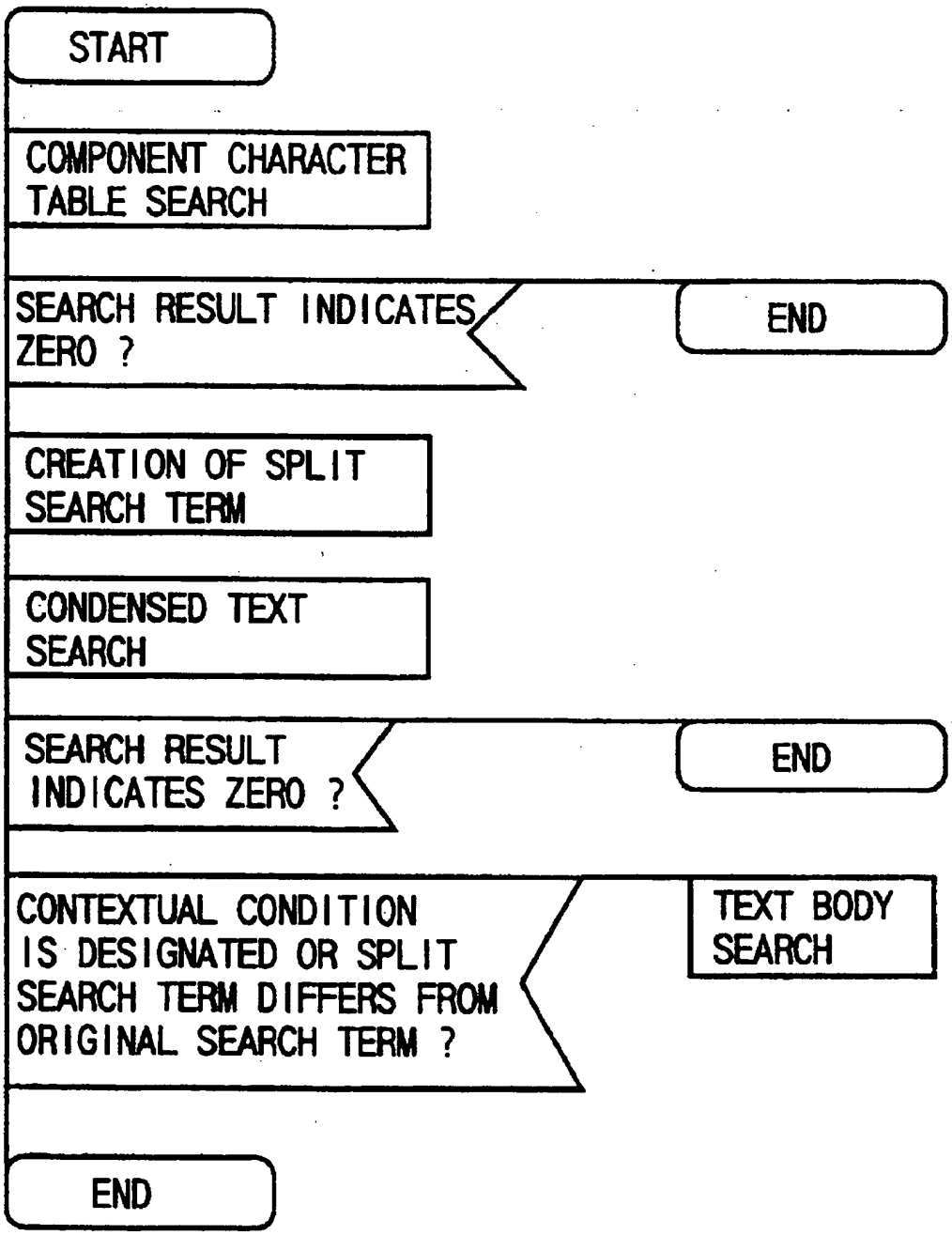

START
COMPONENT CHARACTER TABLE SEARCH
SEARCH RESULT INDICATES ZERO ?
END
CREATION OF SPLIT SEARCH TERM
CONDENSED TEXT SEARCH
SEARCH RESULT INDICATES ZERO ?
END
CONTEXTUAL CONDITION IS DESIGNATED OR SPLIT SEARCH TERM DIFFERS FROM ORIGINAL SEARCH TERM ?
TEXT BODY SEARCH
END

FIG. 15

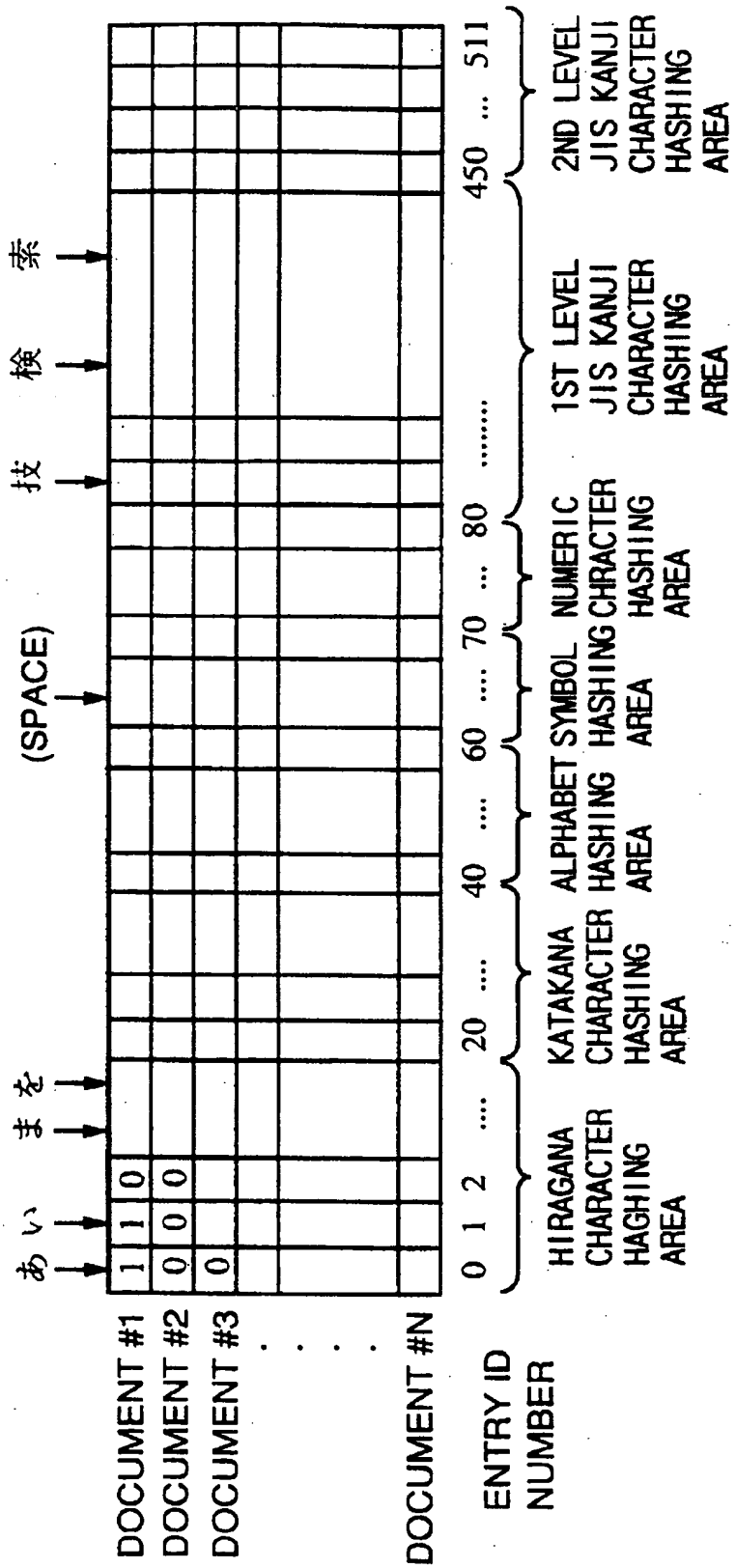

あ い ま を
(SPACE)
技 検 索
DOCUMENT #1
DOCUMENT #2
DOCUMENT #3
DOCUMENT #N
ENTRY ID NUMBER
0 1 2 .... 20 .... 40 .... 60 .... 70 ... 80 ..... 450 ... 511
HIRAGANA CHARACTER HAGHING AREA
KATAKANA CHARACTER HASHING AREA
ALPHABET HASHING AREA
SYMBOL HASHING AREA
NUMERIC CHRACTER HASHING AREA
1ST LEVEL JIS KANJI CHARACTER HASHING AREA
2ND LEVEL JIS KANJI CHARACTER HASHING AREA

FIG. 16

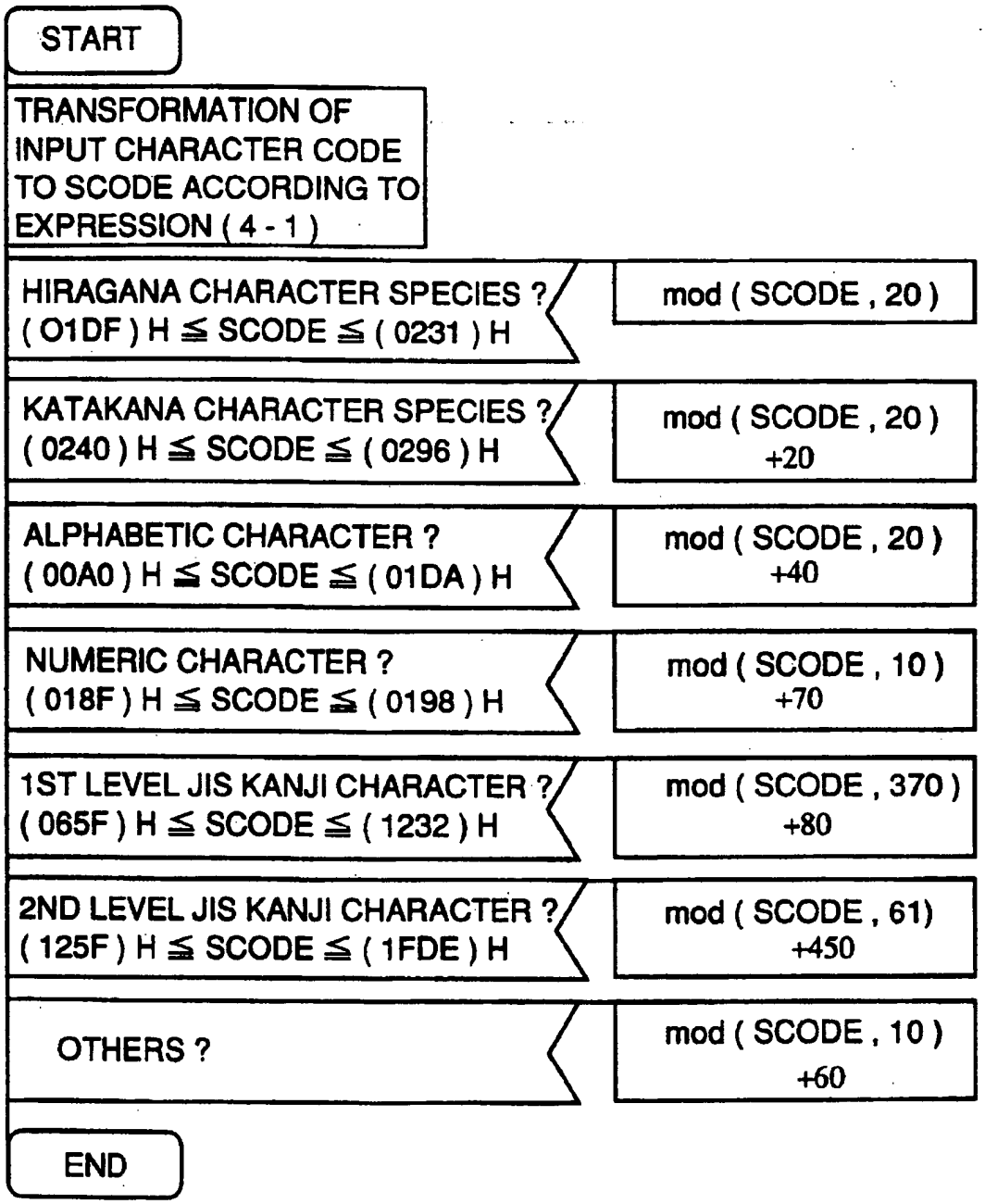

START
TRANSFORMATION OF INPUT CHARACTER CODE TO SCODE ACCORDING TO EXPRESSION ( 4 - 1 )
HIRAGANA CHARACTER SPECIES ? ( O1DF ) H ≦ SCODE ≦ ( 0231 ) H
mod ( SCODE , 20 )
KATAKANA CHARACTER SPECIES ? ( 0240 ) H ≦ SCODE ≦ ( 0296 ) H
mod ( SCODE , 20 ) +20
ALPHABETIC CHARACTER ? ( 00A0 ) H ≦ SCODE ≦ ( 01DA ) H
mod ( SCODE , 20 ) +40
NUMERIC CHARACTER ? ( 018F ) H ≦ SCODE ≦ ( 0198 ) H
mod ( SCODE , 10 ) +70
1ST LEVEL JIS KANJI CHARACTER ? ( 065F ) H ≦ SCODE ≦ ( 1232 ) H
mod ( SCODE , 370 ) +80
2ND LEVEL JIS KANJI CHARACTER ? ( 125F ) H ≦ SCODE ≦ ( 1FDE ) H
mod ( SCODE , 61) +450
OTHERS ?
mod ( SCODE , 10 ) +60
END

FIG. 17

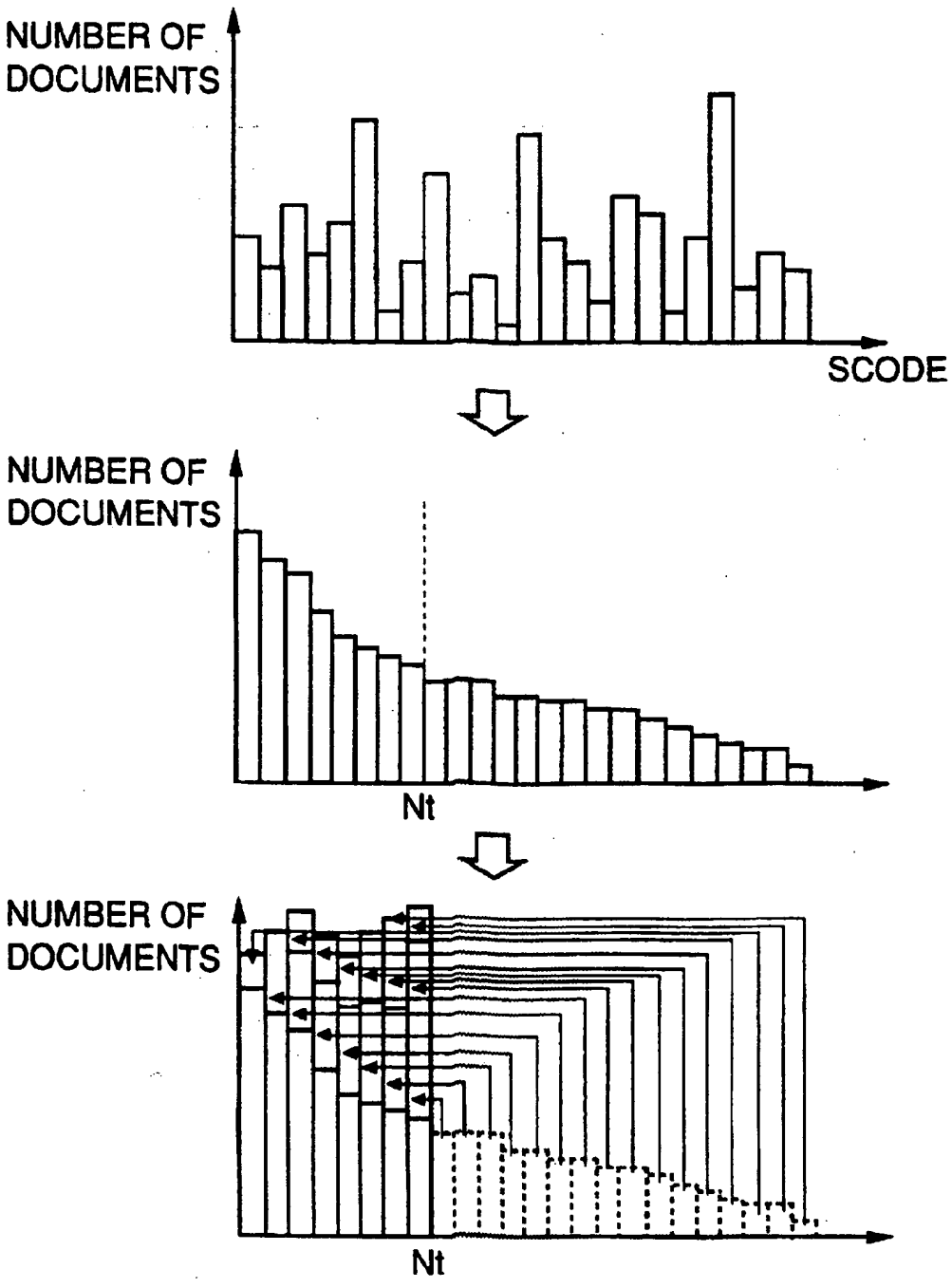

NUMBER OF DOCUMENTS
SCODE
NUMBER OF DOCUMENTS
Nt
NUMBER OF DOCUMENTS
Nt

FIG. 18

検・・・

| SCODE | ENTRY ID NUMBER |
|---|---|
| • | • |
| • | • |
| • | • |
| (095D)H | 150 |
| (095E)H | 356 |
| (095F)H | 231 |
| (0960)H | 483 |
| (0961)H | 2 |
| (0962)H | 256 |
| (0963)H | 25 |
| (0964)H | 67 |
| • | • |
| • | • |
| • | • |

FIG. 19

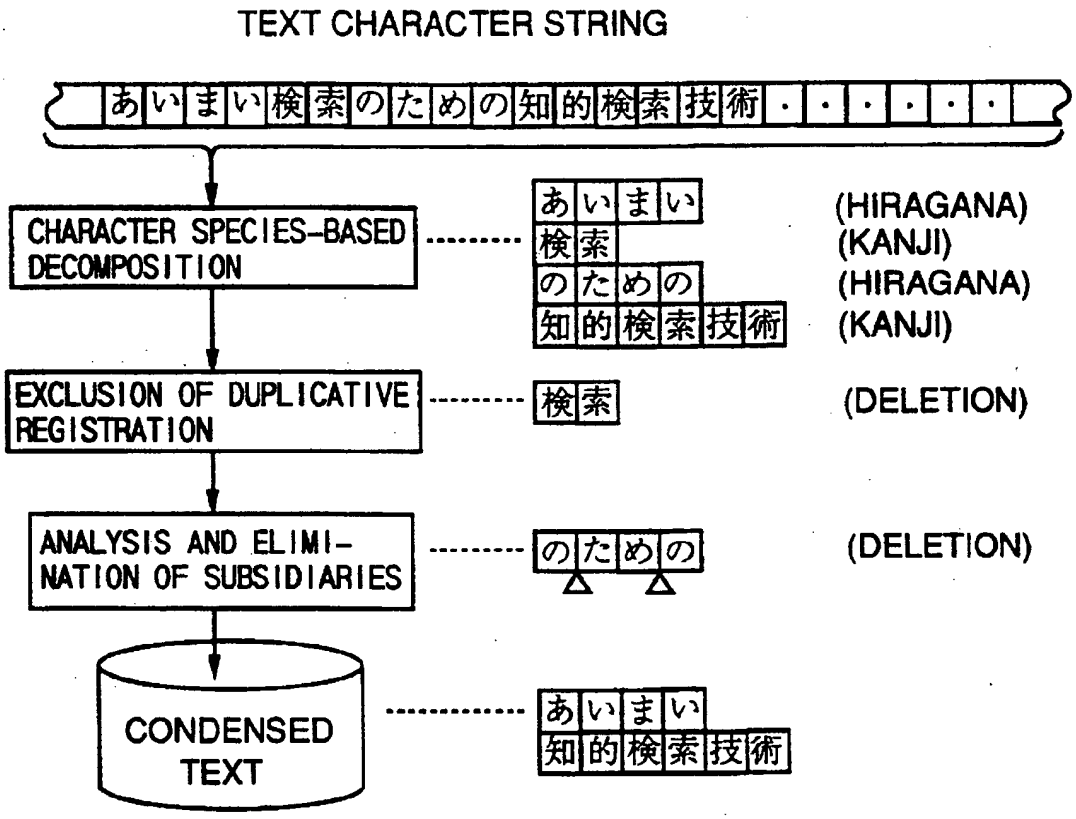
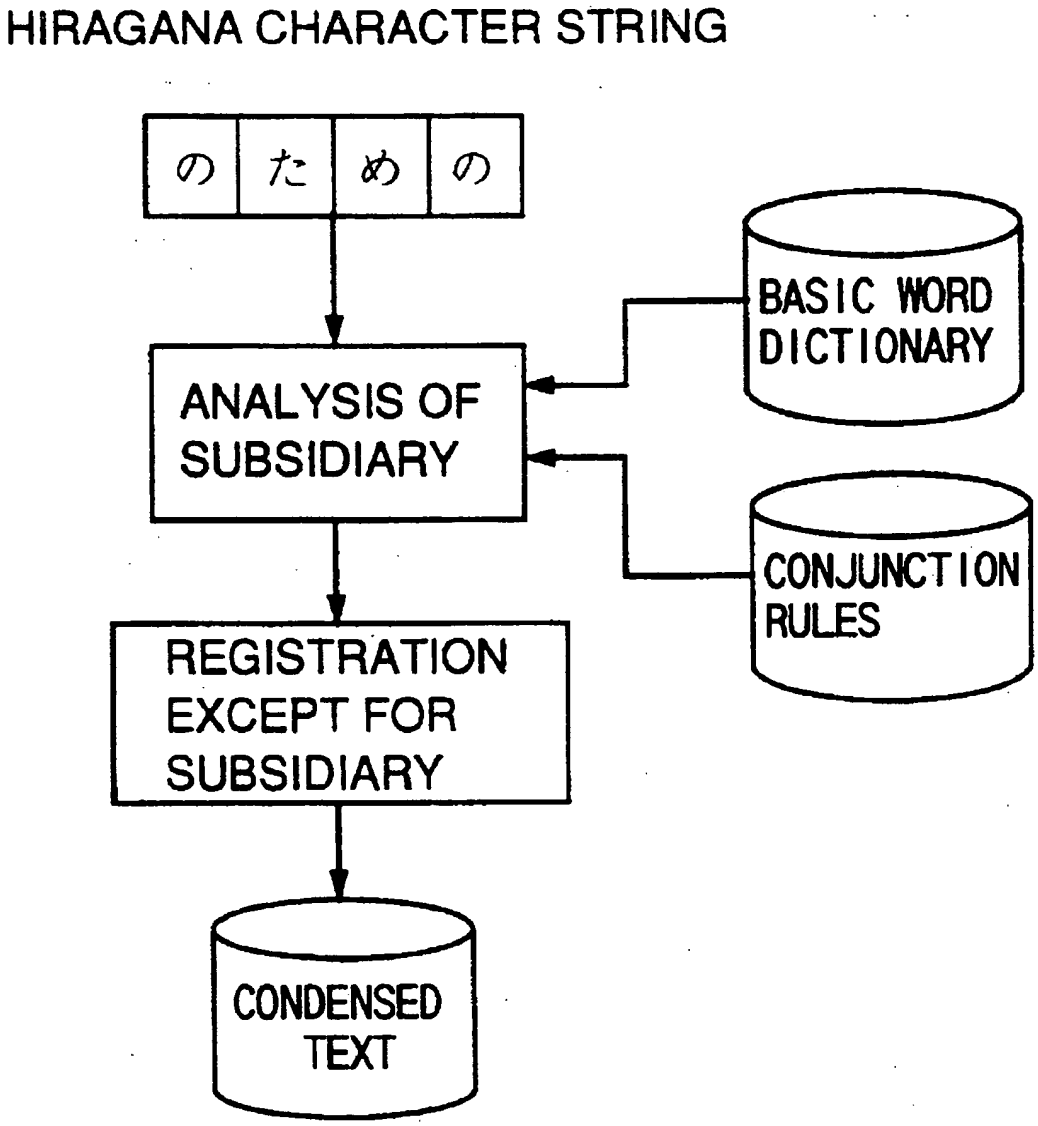

TEXT CHARACTER STRING
あいまい検索のための知的検索技術・・・・・・
CHARACTER SPECIES-BASED DECOMPOSITION
あいまい (HIRAGANA)
検索 (KANJI)
のための (HIRAGANA)
知的検索技術 (KANJI)
EXCLUSION OF DUPLICATIVE REGISTRATION
検索 (DELETION)
ANALYSIS AND ELIMINATION OF SUBSIDIARIES
のため (DELETION)
CONDENSED TEXT
あいまい
知的検索技術

FIG. 20


HIRAGANA CHARACTER STRING
の
た
め
の
BASIC WORD DICTIONARY
ANALYSIS OF SUBSIDIARY
CONJUNCTION RULES
REGISTRATION EXCEPT FOR SUBSIDIARY
CONDENSED TEXT

# FIG. 21

BASIC WORD DICTIONARY

| | |
|---|---|
| 〈 PARTICIPIAL FORM OF VERB “ある” 〉 | : ある |
| 〈 SUBJUNCTIVE FORM OF VERB “ある” 〉 | : あれ |
| 〈 UNENDED FORM OF VERB “なる” 〉 | : なら |
| 〈 CONTINUATIVE FORM OF VERB “なる” 〉 | : なり |
| 〈 UNENDED FORM OF VERB “もつ” 〉 | : もた |
| 〈POSTPOSITIONAL WORD “が” 〉 | : が |
| 〈NOUN “こと” 〉 | : こと |
| 〈NOUN “ため” 〉 | : ため |
| 〈NOUN “の” 〉 | : の |

# FIG. 22

CONJUNCTION RULES

CONJUNCTION RULE 1 : 〈PARTICIPIAL FORM OF VERB “ある”〉+〈NOUN “こと”〉
CONJUNCTION RULE 2 : 〈PARTICIPIAL FORM OF VERB “もつ”〉+〈NOUN “ため”〉
CONJUNCTION RULE 3 : 〈NOUN “こと”〉+〈POSTPOSICNAL WORD “が”〉
CONJUNCTION RULE 5 : 〈PARTICIPIAL FORM OF VERB “する”〉+〈NOUN “こと”〉
CONJUNCTION RULE 6 : 〈POSTPOSITIONAL WORD “の”〉+〈NOUN “ため”〉
CONJUNCTION RULE 7 : 〈NOUN “ため”〉+〈POSTPOSICNAL WORD “の”〉
CONJUNCTION RULE 8 : 〈CONJUNCTION〉+〈POSTPOSICNAL WORD “は”〉
CONJUNCTION RULE 9 : 〈UNENDED FORM OF VERB “する”〉+〈POSTPOSICNAL WORD “は”〉
CONJUNCTION RULE 10 : 〈NOUN “こと”〉+〈POSTPOSICNAL WORD “に”〉
CONJUNCTION RULE 11 : 〈POSTPOSITIONAL WORD “で”〉+〈ENDED FORM OF VERB “ある”〉
CONJUNCTION RULE 12 : 〈UNENDED FORM OF VERB “ある”〉+〈POSTPOSICNAL WORD “ば”〉
CONJUNCTION RULE 13 : 〈POSTPOSITIONAL WORD “ある”〉+〈ENDED FORM OF VERB “だ”〉

FIG. 23

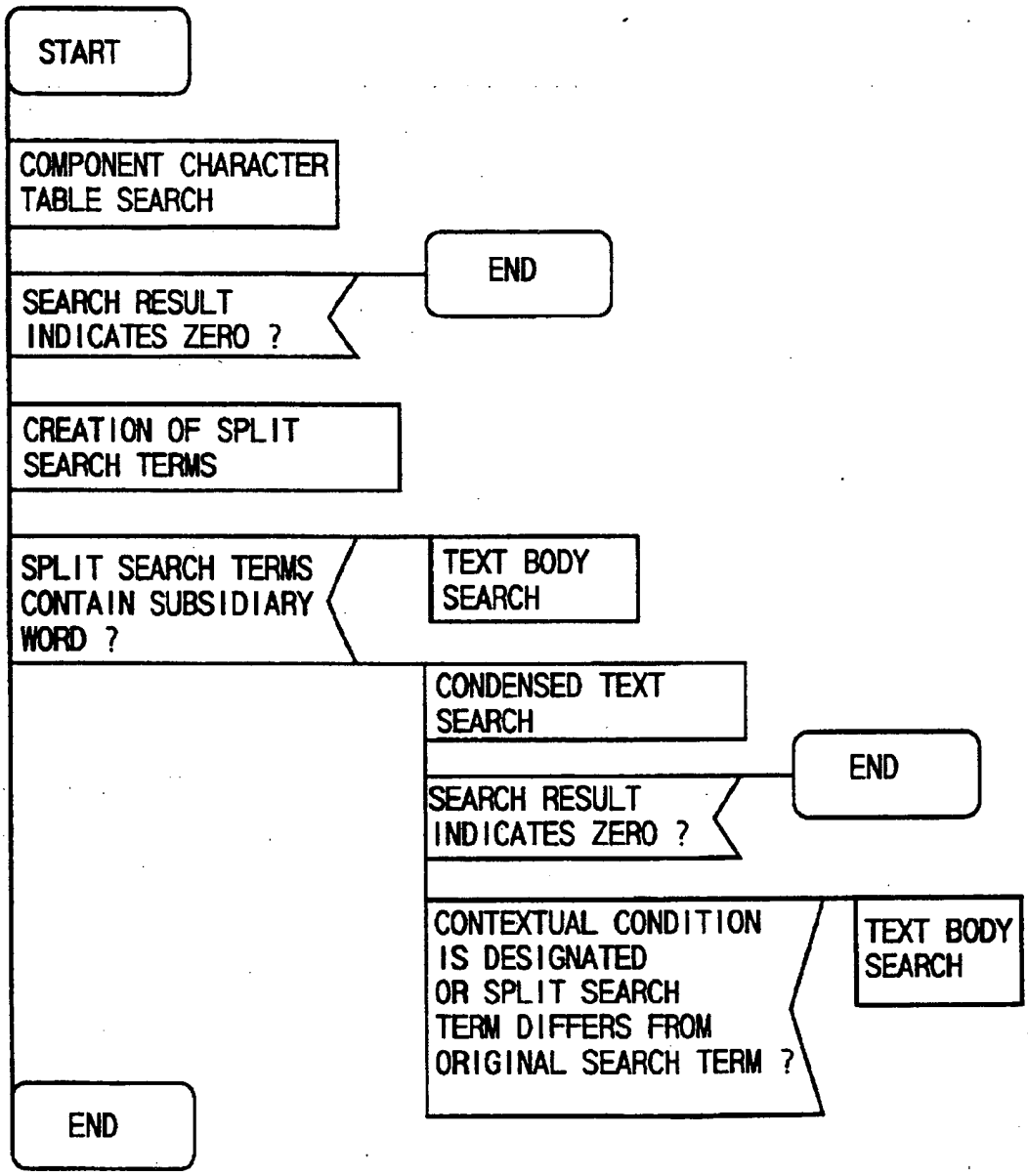

START
COMPONENT CHARACTER TABLE SEARCH
SEARCH RESULT INDICATES ZERO ?
END
CREATION OF SPLIT SEARCH TERMS
SPLIT SEARCH TERMS CONTAIN SUBSIDIARY WORD ?
TEXT BODY SEARCH
CONDENSED TEXT SEARCH
SEARCH RESULT INDICATES ZERO ?
END
CONTEXTUAL CONDITION IS DESIGNATED OR SPLIT SEARCH TERM DIFFERS FROM ORIGINAL SEARCH TERM ?
TEXT BODY SEARCH
END

FIG. 24

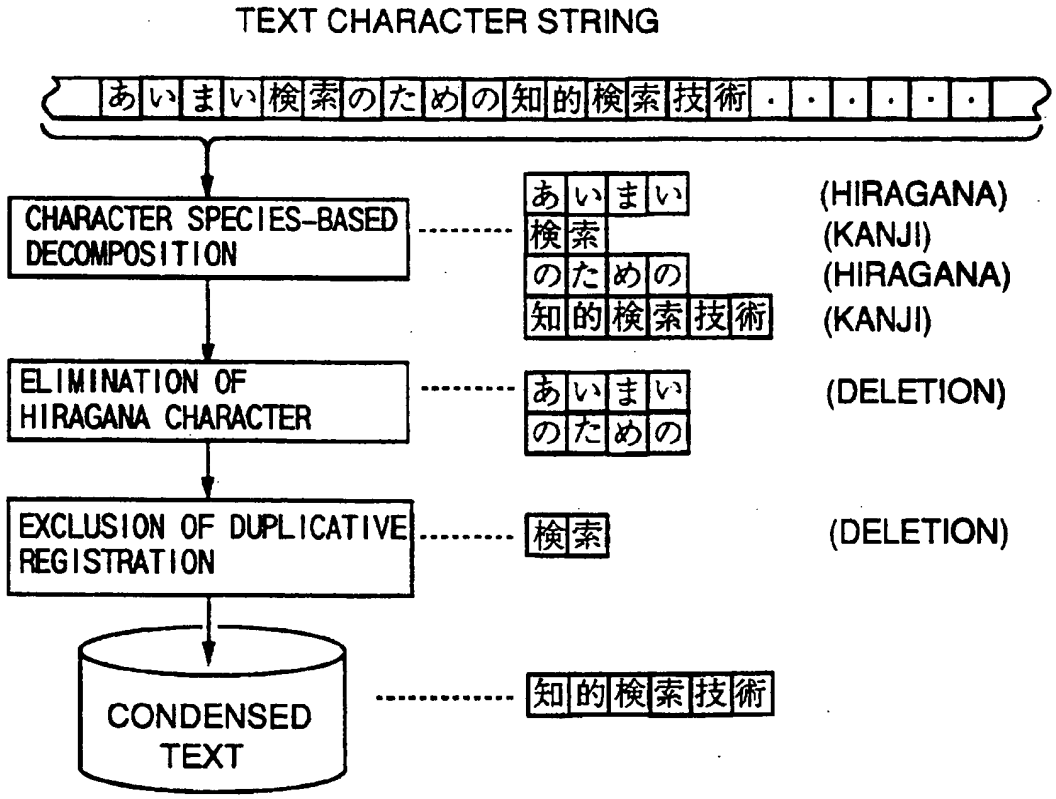

TEXT CHARACTER STRING
あいまい検索のための知的検索技術・・・・・・
CHARACTER SPECIES-BASED DECOMPOSITION
あいまい (HIRAGANA)
検索 (KANJI)
のための (HIRAGANA)
知的検索技術 (KANJI)
ELIMINATION OF HIRAGANA CHARACTER
あいまい (DELETION)
のための
EXCLUSION OF DUPLICATIVE REGISTRATION
検索 (DELETION)
CONDENSED TEXT
知的検索技術

FIG. 25

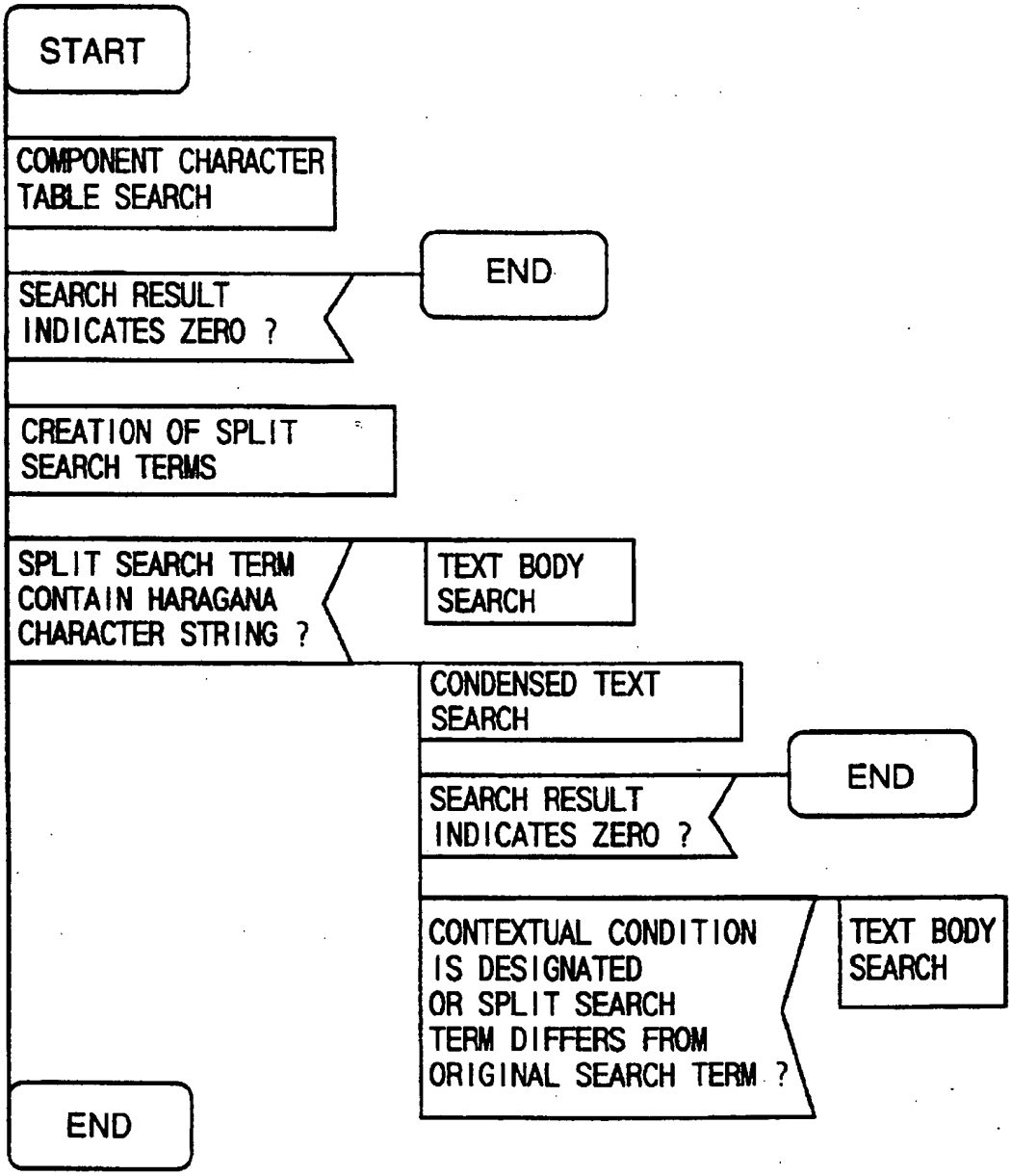

START
COMPONENT CHARACTER TABLE SEARCH
SEARCH RESULT INDICATES ZERO ?
END
CREATION OF SPLIT SEARCH TERMS
SPLIT SEARCH TERM CONTAIN HARAGANA CHARACTER STRING ?
TEXT BODY SEARCH
CONDENSED TEXT SEARCH
SEARCH RESULT INDICATES ZERO ?
END
CONTEXTUAL CONDITION IS DESIGNATED OR SPLIT SEARCH TERM DIFFERS FROM ORIGINAL SEARCH TERM ?
TEXT BODY SEARCH
END

FIG. 26

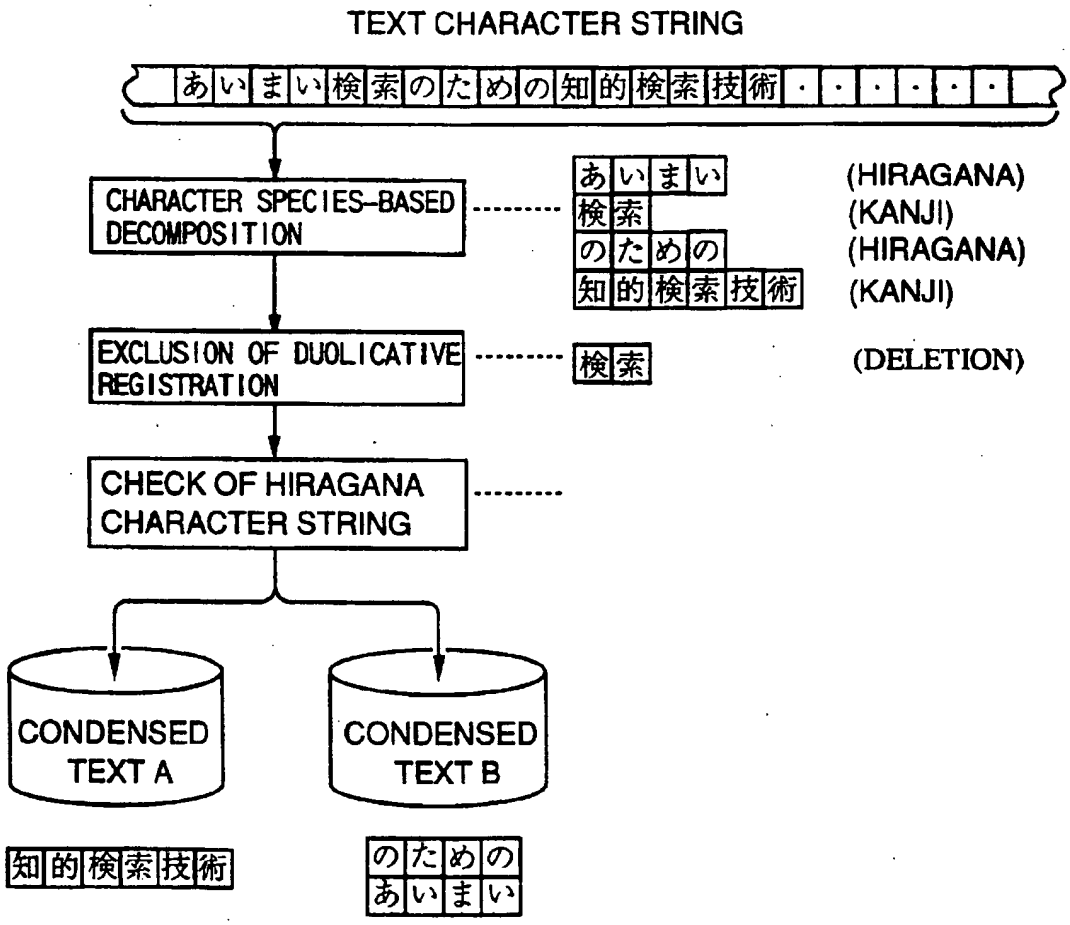

TEXT CHARACTER STRING
あいまい検索のための知的検索技術
CHARACTER SPECIES-BASED DECOMPOSITION
あいまい (HIRAGANA)
検索 (KANJI)
のための (HIRAGANA)
知的検索技術 (KANJI)
EXCLUSION OF DUOLICATIVE REGISTRATION
検索 (DELETION)
CHECK OF HIRAGANA CHARACTER STRING
CONDENSED TEXT A
CONDENSED TEXT B
知的検索技術
のためのあいまい

FIG. 27

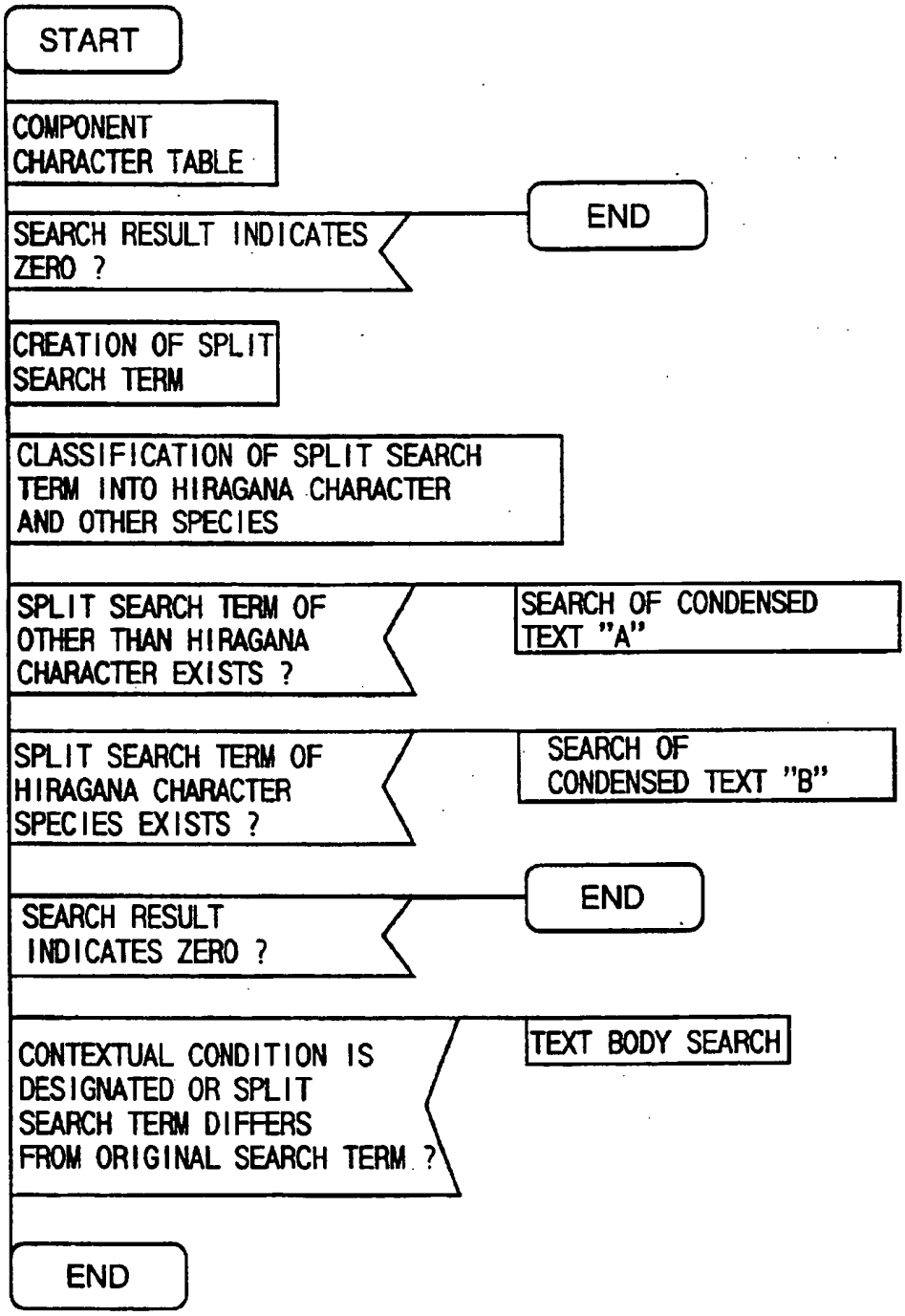

START
COMPONENT CHARACTER TABLE
SEARCH RESULT INDICATES ZERO ?
END
CREATION OF SPLIT SEARCH TERM
CLASSIFICATION OF SPLIT SEARCH TERM INTO HIRAGANA CHARACTER AND OTHER SPECIES
SPLIT SEARCH TERM OF OTHER THAN HIRAGANA CHARACTER EXISTS ?
SEARCH OF CONDENSED TEXT "A"
SPLIT SEARCH TERM OF HIRAGANA CHARACTER SPECIES EXISTS ?
SEARCH OF CONDENSED TEXT "B"
SEARCH RESULT INDICATES ZERO ?
END
CONTEXTUAL CONDITION IS DESIGNATED OR SPLIT SEARCH TERM DIFFERS FROM ORIGINAL SEARCH TERM ?
TEXT BODY SEARCH
END

FIG. 28

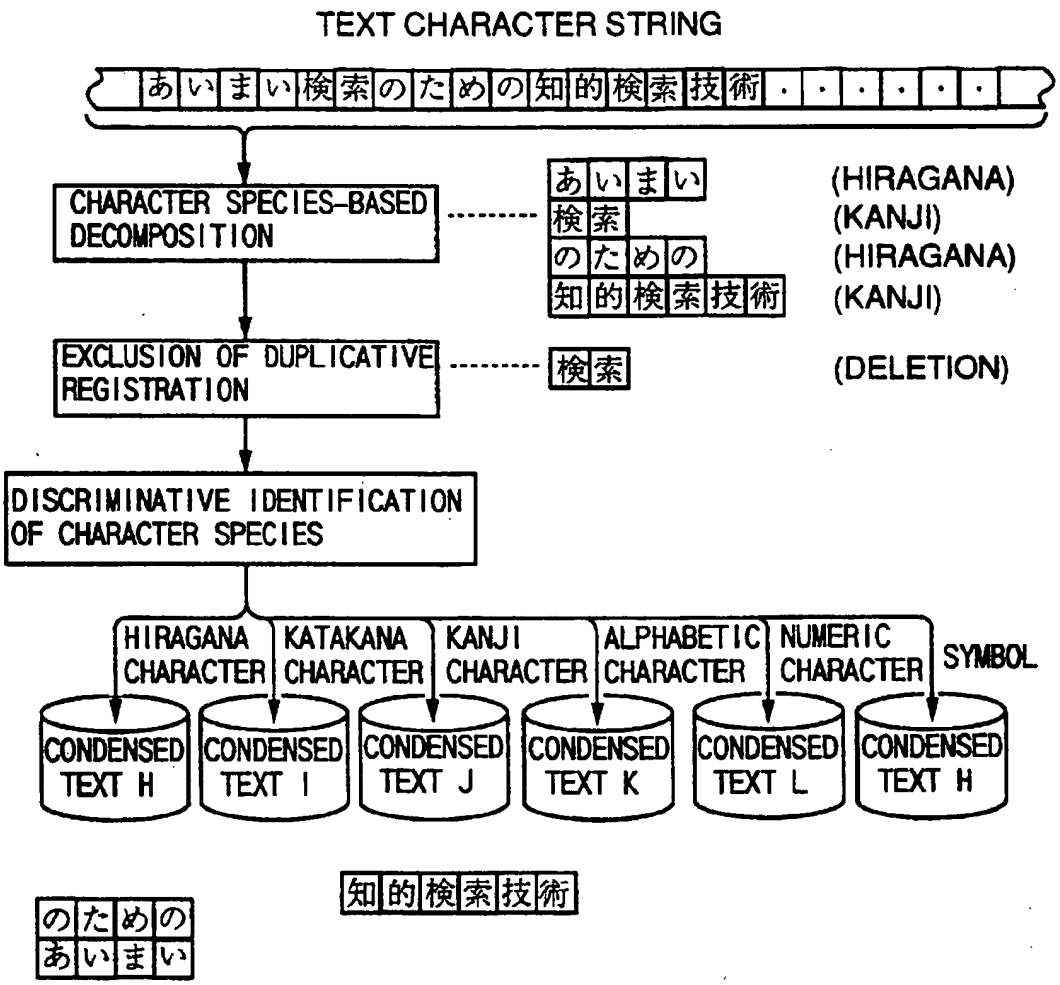

TEXT CHARACTER STRING
あいまい検索のための知的検索技術
CHARACTER SPECIES-BASED DECOMPOSITION
あいまい (HIRAGANA)
検索 (KANJI)
のための (HIRAGANA)
知的検索技術 (KANJI)
EXCLUSION OF DUPLICATIVE REGISTRATION
検索 (DELETION)
DISCRIMINATIVE IDENTIFICATION OF CHARACTER SPECIES
HIRAGANA CHARACTER
KATAKANA CHARACTER
KANJI CHARACTER
ALPHABETIC CHARACTER
NUMERIC CHARACTER
SYMBOL
CONDENSED TEXT H
CONDENSED TEXT I
CONDENSED TEXT J
CONDENSED TEXT K
CONDENSED TEXT L
CONDENSED TEXT H
知的検索技術
のためのあいまい

FIG. 29

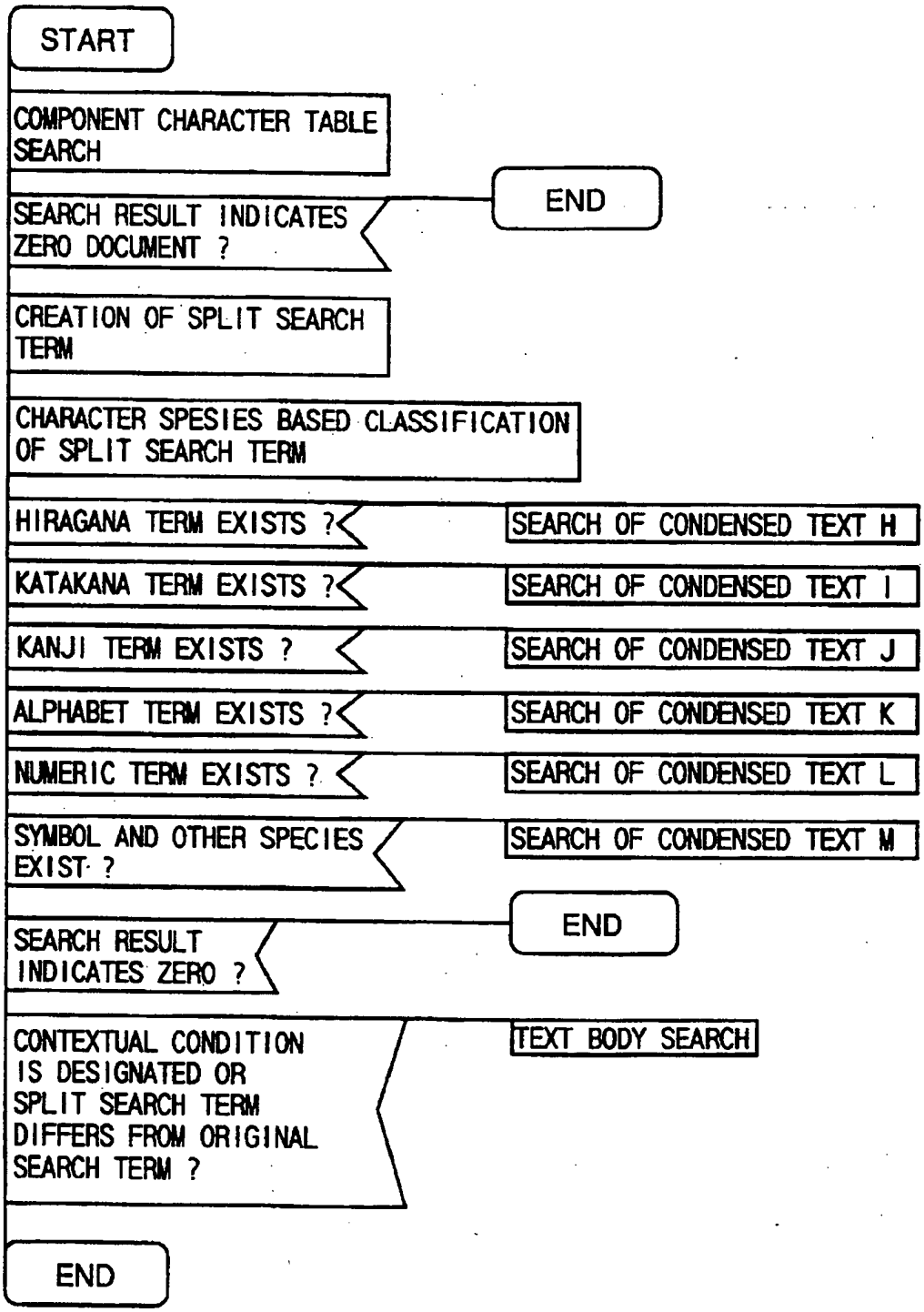

START
COMPONENT CHARACTER TABLE SEARCH
SEARCH RESULT INDICATES ZERO DOCUMENT ?
END
CREATION OF SPLIT SEARCH TERM
CHARACTER SPESIES BASED CLASSIFICATION OF SPLIT SEARCH TERM
HIRAGANA TERM EXISTS ?
SEARCH OF CONDENSED TEXT H
KATAKANA TERM EXISTS ?
SEARCH OF CONDENSED TEXT I
KANJI TERM EXISTS ?
SEARCH OF CONDENSED TEXT J
ALPHABET TERM EXISTS ?
SEARCH OF CONDENSED TEXT K
NUMERIC TERM EXISTS ?
SEARCH OF CONDENSED TEXT L
SYMBOL AND OTHER SPECIES EXIST ?
SEARCH OF CONDENSED TEXT M
END
SEARCH RESULT INDICATES ZERO ?
CONTEXTUAL CONDITION IS DESIGNATED OR SPLIT SEARCH TERM DIFFERS FROM ORIGINAL SEARCH TERM ?
TEXT BODY SEARCH
END

FIG. 30

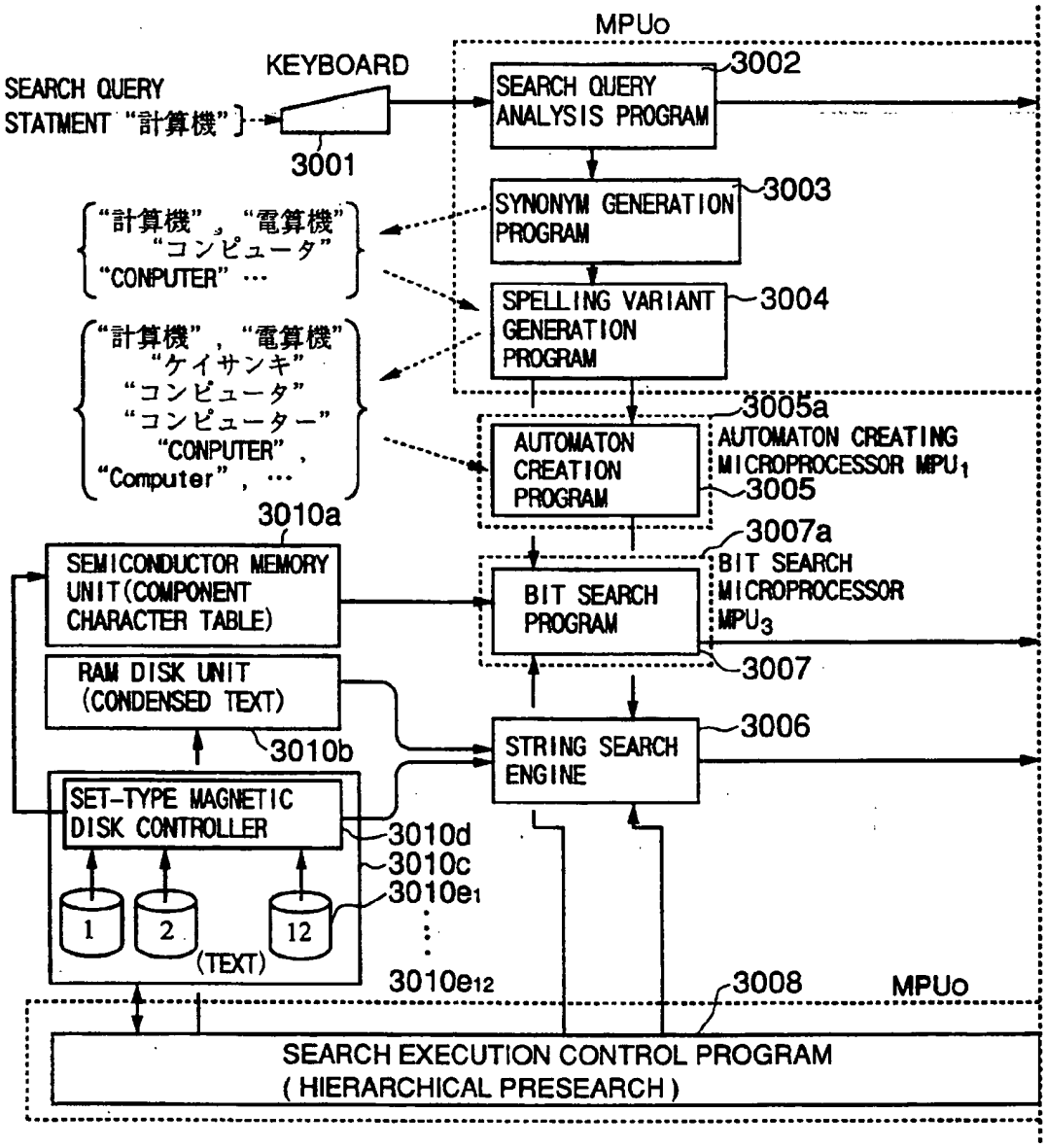

MPUo
KEYBOARD
SEARCH QUERY STATMENT "計算機"
3001
SEARCH QUERY ANALYSIS PROGRAM
3002
SYNONYM GENERATION PROGRAM
3003
"計算機", "電算機" "コンピュータ" "CONPUTER" ···
SPELLING VARIANT GENERATION PROGRAM
3004
"計算機", "電算機" "ケイサンキ" "コンピュータ" "コンピューター" "CONPUTER", "Computer", ···
AUTOMATON CREATION PROGRAM
3005a
AUTOMATON CREATING MICROPROCESSOR MPU1
3005
3010a
SEMICONDUCTOR MEMORY UNIT(COMPONENT CHARACTER TABLE)
BIT SEARCH PROGRAM
3007a
BIT SEARCH MICROPROCESSOR MPU3
3007
RAM DISK UNIT (CONDENSED TEXT)
3010b
STRING SEARCH ENGINE
3006
SET-TYPE MAGNETIC DISK CONTROLLER
3010d
3010c
3010e1
1
2
12
(TEXT)
3010e12
3008
MPUo
SEARCH EXECUTION CONTROL PROGRAM (HIERARCHICAL PRESEARCH)

FIG. 31

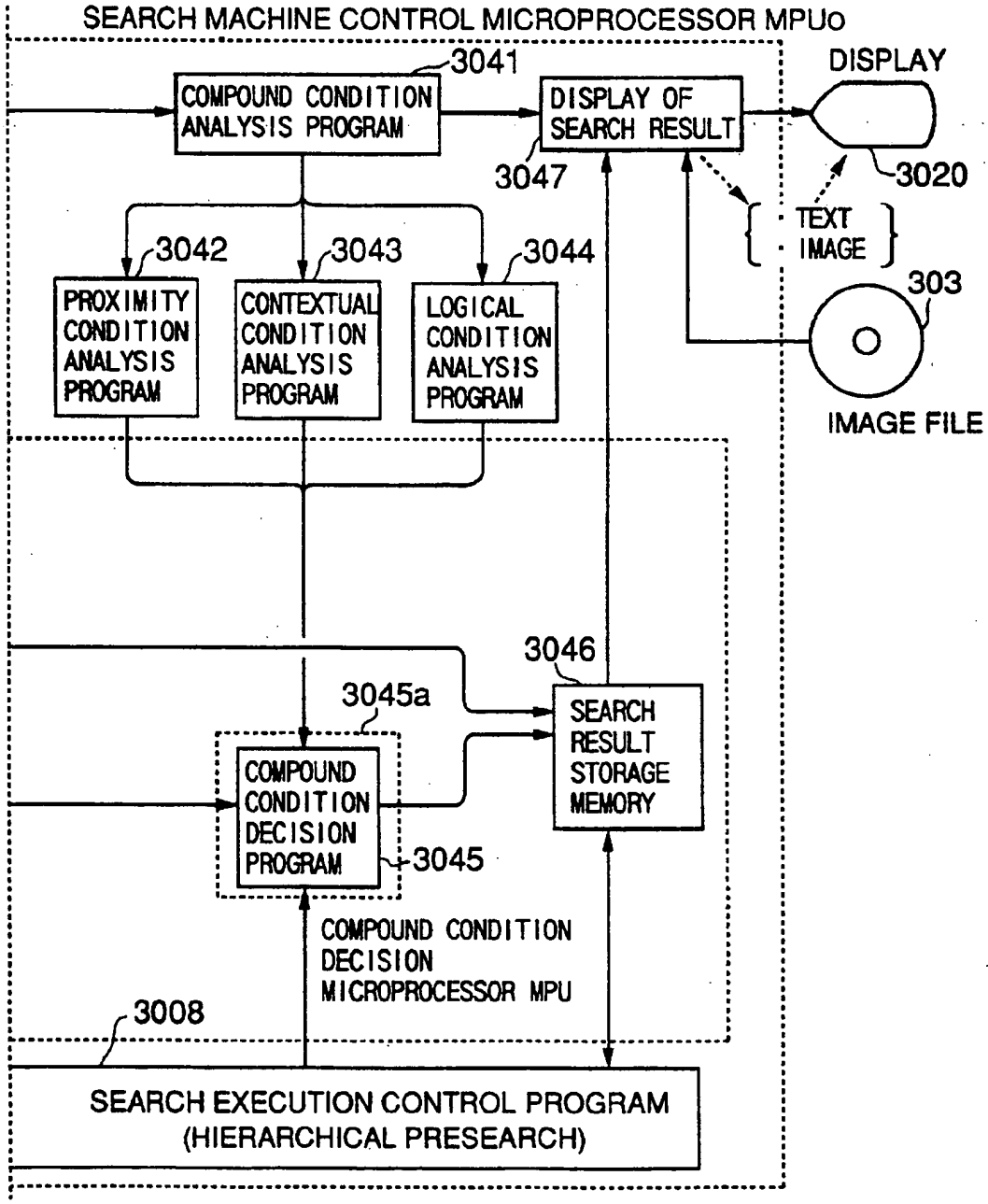

SEARCH MACHINE CONTROL MICROPROCESSOR MPUo
3041
COMPOUND CONDITION ANALYSIS PROGRAM
DISPLAY OF SEARCH RESULT
DISPLAY
3047
3020
TEXT IMAGE
3042
3043
3044
PROXIMITY CONDITION ANALYSIS PROGRAM
CONTEXTUAL CONDITION ANALYSIS PROGRAM
LOGICAL CONDITION ANALYSIS PROGRAM
303
IMAGE FILE
3046
3045a
SEARCH RESULT STORAGE MEMORY
COMPOUND CONDITION DECISION PROGRAM
3045
COMPOUND CONDITION DECISION MICROPROCESSOR MPU
3008
SEARCH EXECUTION CONTROL PROGRAM (HIERARCHICAL PRESEARCH)

FIG. 32

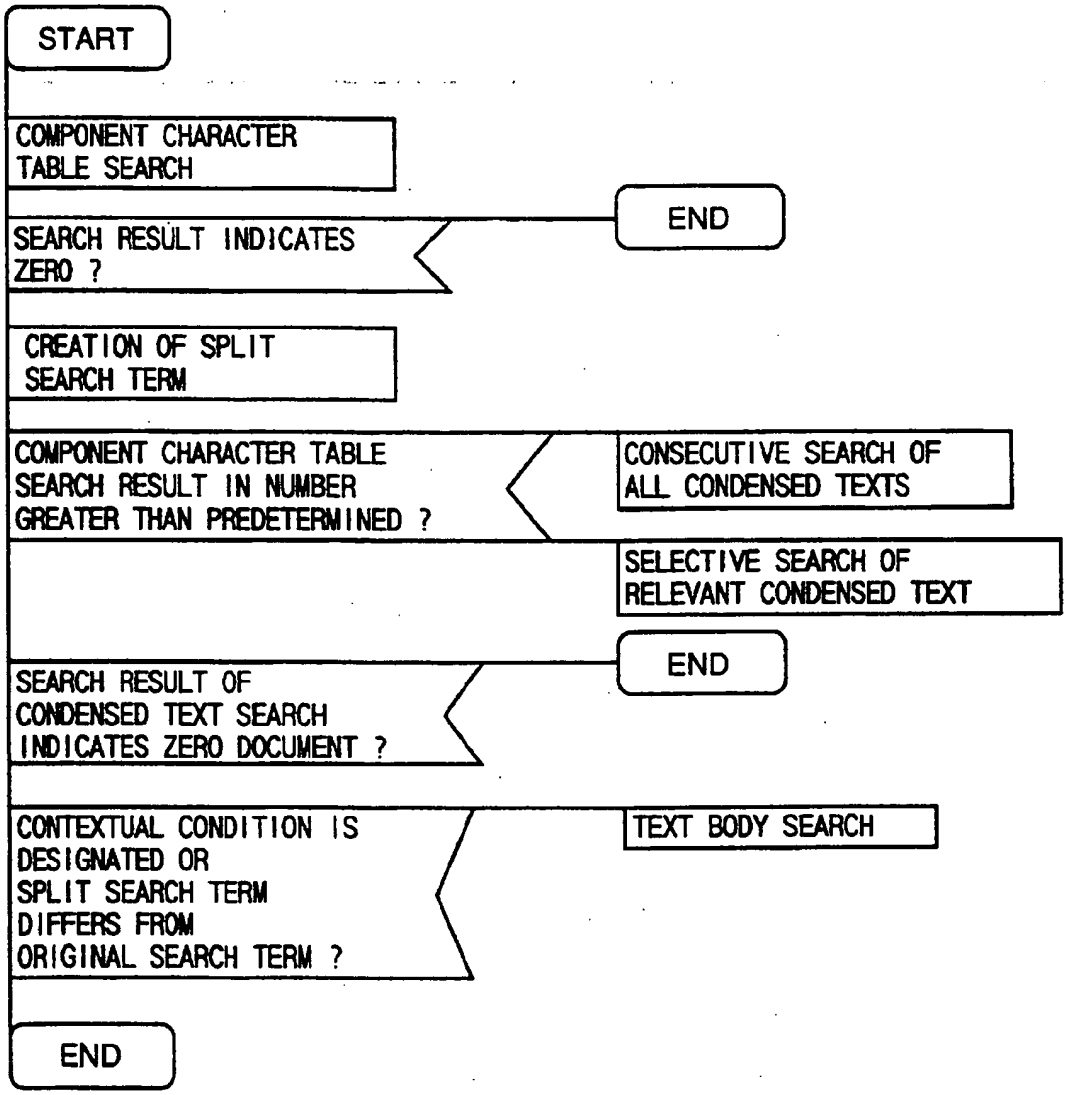

START
COMPONENT CHARACTER TABLE SEARCH
SEARCH RESULT INDICATES ZERO ?
END
CREATION OF SPLIT SEARCH TERM
COMPONENT CHARACTER TABLE SEARCH RESULT IN NUMBER GREATER THAN PREDETERMINED ?
CONSECUTIVE SEARCH OF ALL CONDENSED TEXTS
SELECTIVE SEARCH OF RELEVANT CONDENSED TEXT
END
SEARCH RESULT OF CONDENSED TEXT SEARCH INDICATES ZERO DOCUMENT ?
CONTEXTUAL CONDITION IS DESIGNATED OR SPLIT SEARCH TERM DIFFERS FROM ORIGINAL SEARCH TERM ?
TEXT BODY SEARCH
END

FIG. 33

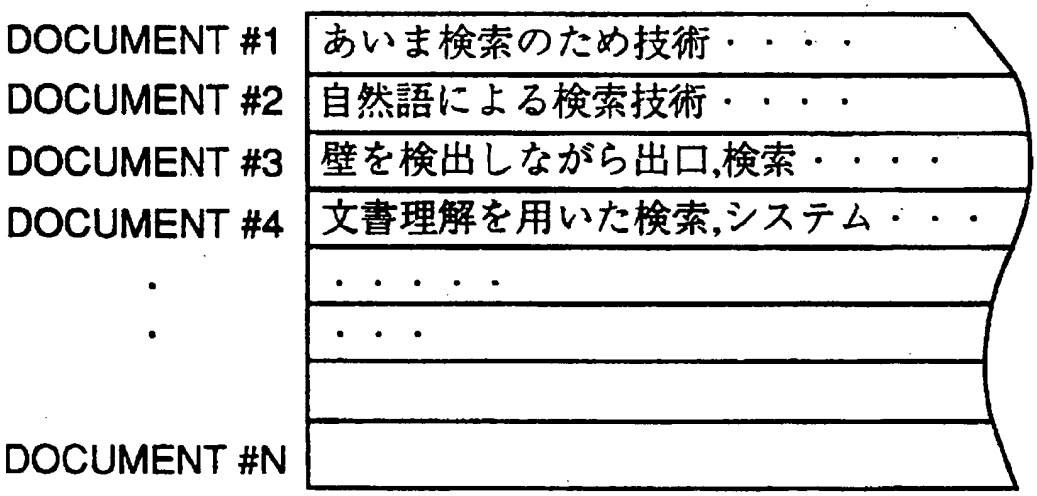

DOCUMENT #1
あいま検索のため技術・・・・
DOCUMENT #2
自然語による検索技術・・・・
DOCUMENT #3
壁を検出しながら出口,検索・・・・
DOCUMENT #4
文書理解を用いた検索,システム・・・
・・・・・・
・・・
DOCUMENT #N

FIG. 34

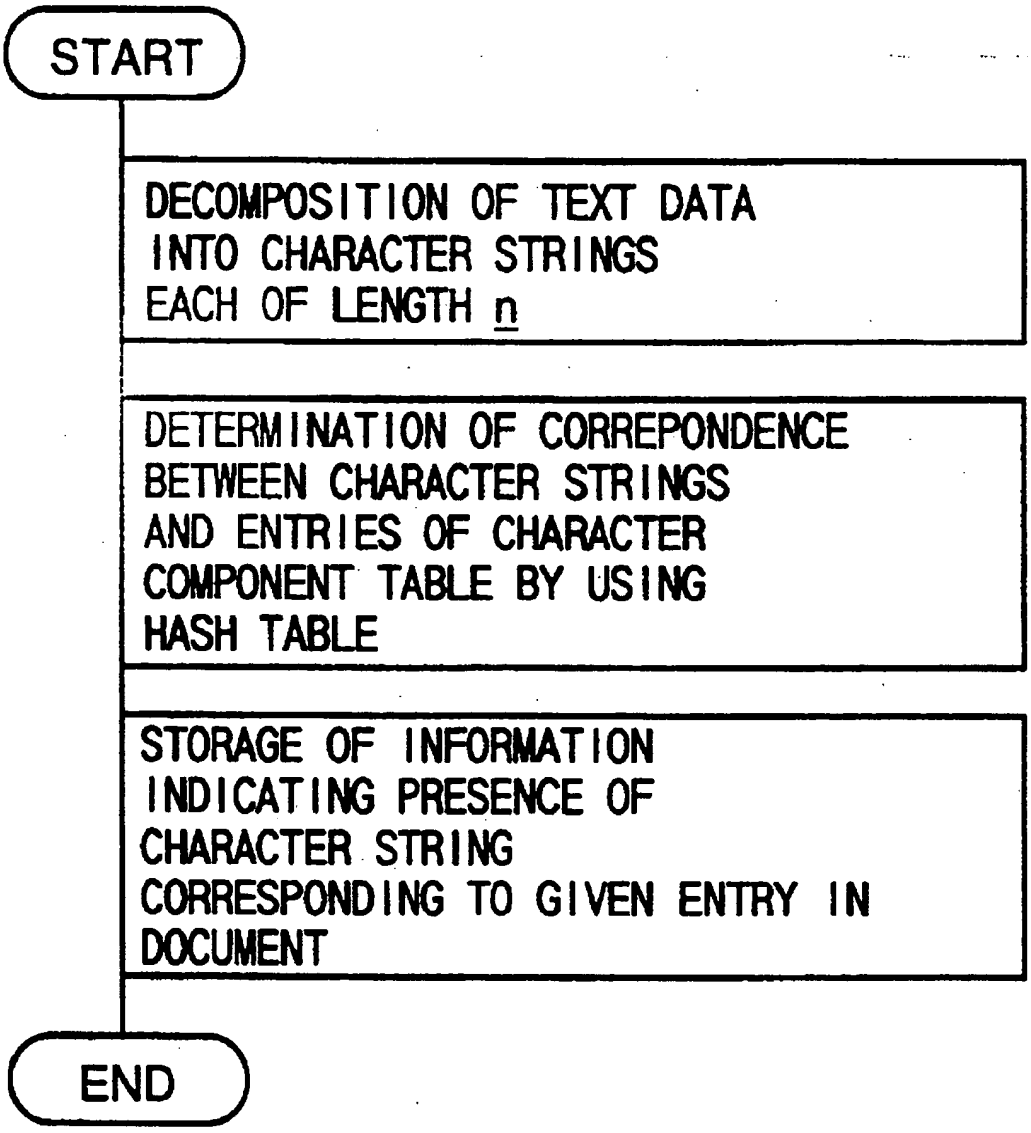

START
DECOMPOSITION OF TEXT DATA INTO CHARACTER STRINGS EACH OF LENGTH n
DETERMINATION OF CORREPONDENCE BETWEEN CHARACTER STRINGS AND ENTRIES OF CHARACTER COMPONENT TABLE BY USING HASH TABLE
STORAGE OF INFORMATION INDICATING PRESENCE OF CHARACTER STRING CORRESPONDING TO GIVEN ENTRY IN DOCUMENT
END

FIG. 35

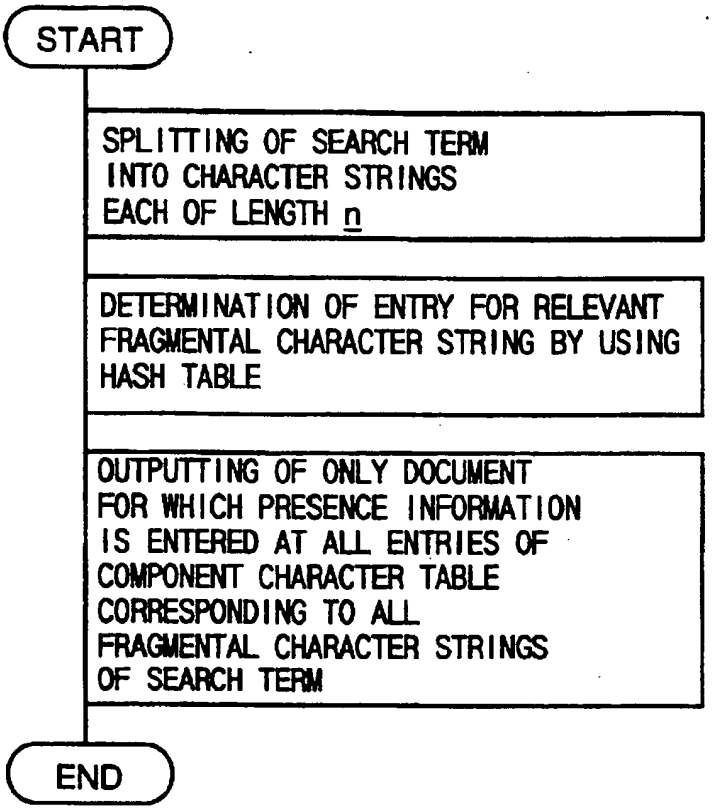

START
SPLITTING OF SEARCH TERM INTO CHARACTER STRINGS EACH OF LENGTH n
DETERMINATION OF ENTRY FOR RELEVANT FRAGMENTAL CHARACTER STRING BY USING HASH TABLE
OUTPUTTING OF ONLY DOCUMENT FOR WHICH PRESENCE INFORMATION IS ENTERED AT ALL ENTRIES OF COMPONENT CHARACTER TABLE CORRESPONDING TO ALL FRAGMENTAL CHARACTER STRINGS OF SEARCH TERM
END

FIG. 36

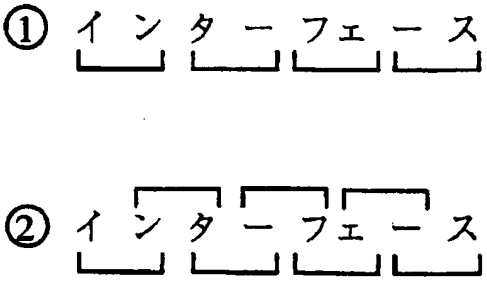

① イ ン タ ー フェ ー ス
② イ ン タ ー フェ ー ス

FIG. 37

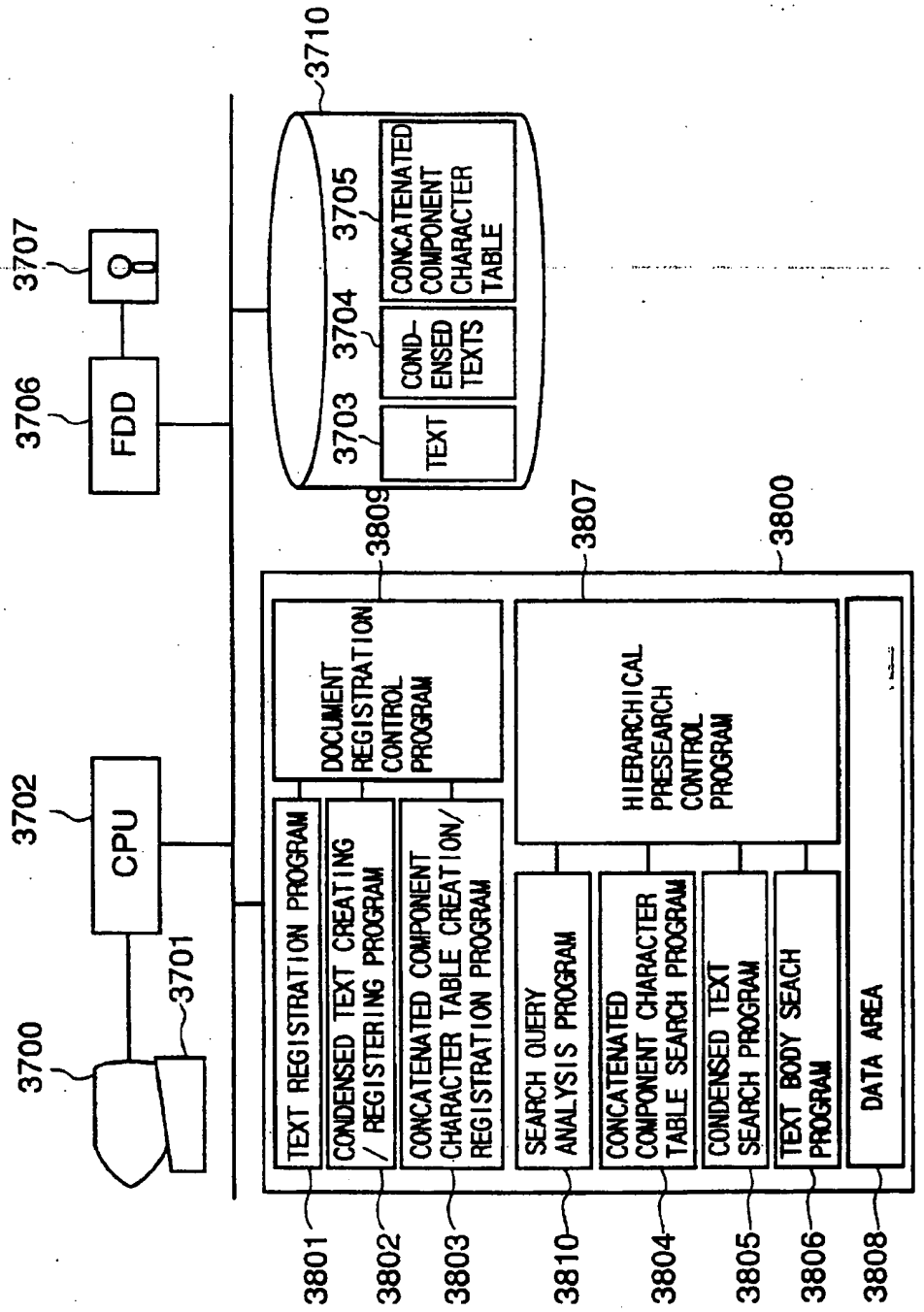

3710
3707
3706
FDD
3705
CONCATENATED COMPONENT CHARACTER TABLE
3704
COND-ENSED TEXTS
3703
TEXT
3809
3807
3800
3702
CPU
3701
3700
DOCUMENT REGISTRATION CONTROL PROGRAM
HIERARCHICAL PRESEARCH CONTROL PROGRAM
TEXT REGISTRATION PROGRAM
CONDENSED TEXT CREATING / REGISTERING PROGRAM
CONCATENATED COMPONENT CHARACTER TABLE CREATION/ REGISTRATION PROGRAM
SEARCH QUERY ANALYSIS PROGRAM
CONCATENATED COMPONENT CHARACTER TABLE SEARCH PROGRAM
CONDENSED TEXT SEARCH PROGRAM
TEXT BODY SEACH PROGRAM
DATA AREA
3801
3802
3803
3810
3804
3805
3806
3808

FIG. 38

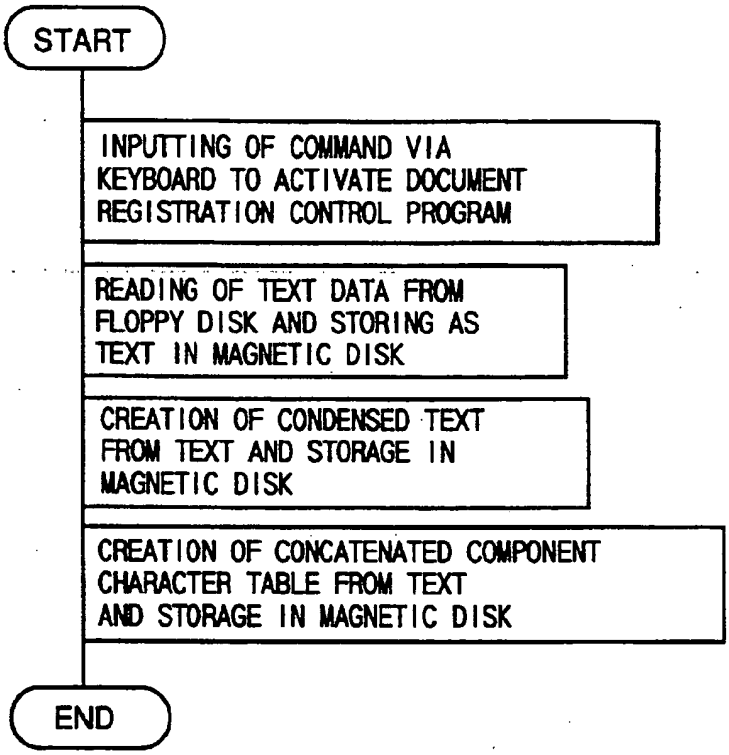

START
INPUTTING OF COMMAND VIA KEYBOARD TO ACTIVATE DOCUMENT REGISTRATION CONTROL PROGRAM
READING OF TEXT DATA FROM FLOPPY DISK AND STORING AS TEXT IN MAGNETIC DISK
CREATION OF CONDENSED TEXT FROM TEXT AND STORAGE IN MAGNETIC DISK
CREATION OF CONCATENATED COMPONENT CHARACTER TABLE FROM TEXT AND STORAGE IN MAGNETIC DISK
END

FIG. 39

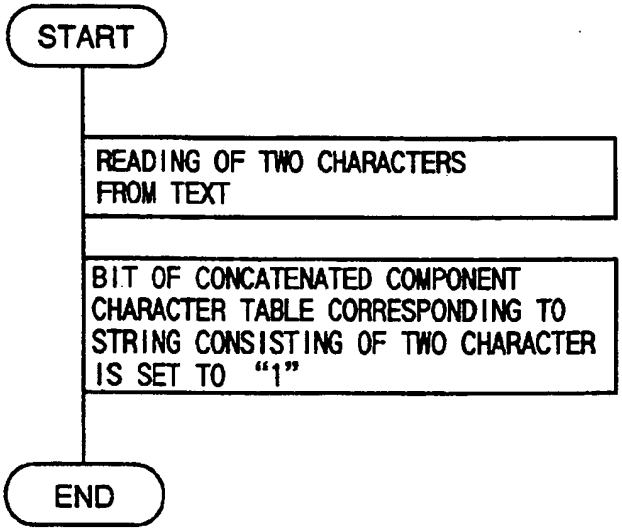

START
READING OF TWO CHARACTERS FROM TEXT
BIT OF CONCATENATED COMPONENT CHARACTER TABLE CORRESPONDING TO STRING CONSISTING OF TWO CHARACTER IS SET TO "1"
END

FIG. 40

| | |
|---|---|
| DOCUMENT #1 | ・・オンラインサービスが経済発展の・・ |
| DOCUMENT #2 | ・・ライオンの生態については未だに・・ |
| DOCUMENT #3 | ・・戦争後のイランとイラクの関係は・・ |

| | ああああい | ・ | オン | イオ | イラ | イン | ライ | ラン | ・ | 関係経済生態戦争 | ・ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| DOCUMENT #1 | 0 0 | | 1 | 0 | 0 | 1 | 1 | 0 | | 0 1 0 0 | |
| DOCUMENT #2 | 0 0 | | 1 | 1 | 0 | 0 | 1 | 0 | | 0 0 1 0 | |
| DOCUMENT #3 | 0 0 | | 0 | 0 | 1 | 0 | 0 | 1 | | 1 0 0 1 | |
| . | | | | | | | | | | | |
| . | | | | | | | | | | | |
| DOCUMENT #N | 0 0 | | 0 | 0 | 0 | 0 | 0 | 0 | | 0 0 0 0 | |

FIG. 41

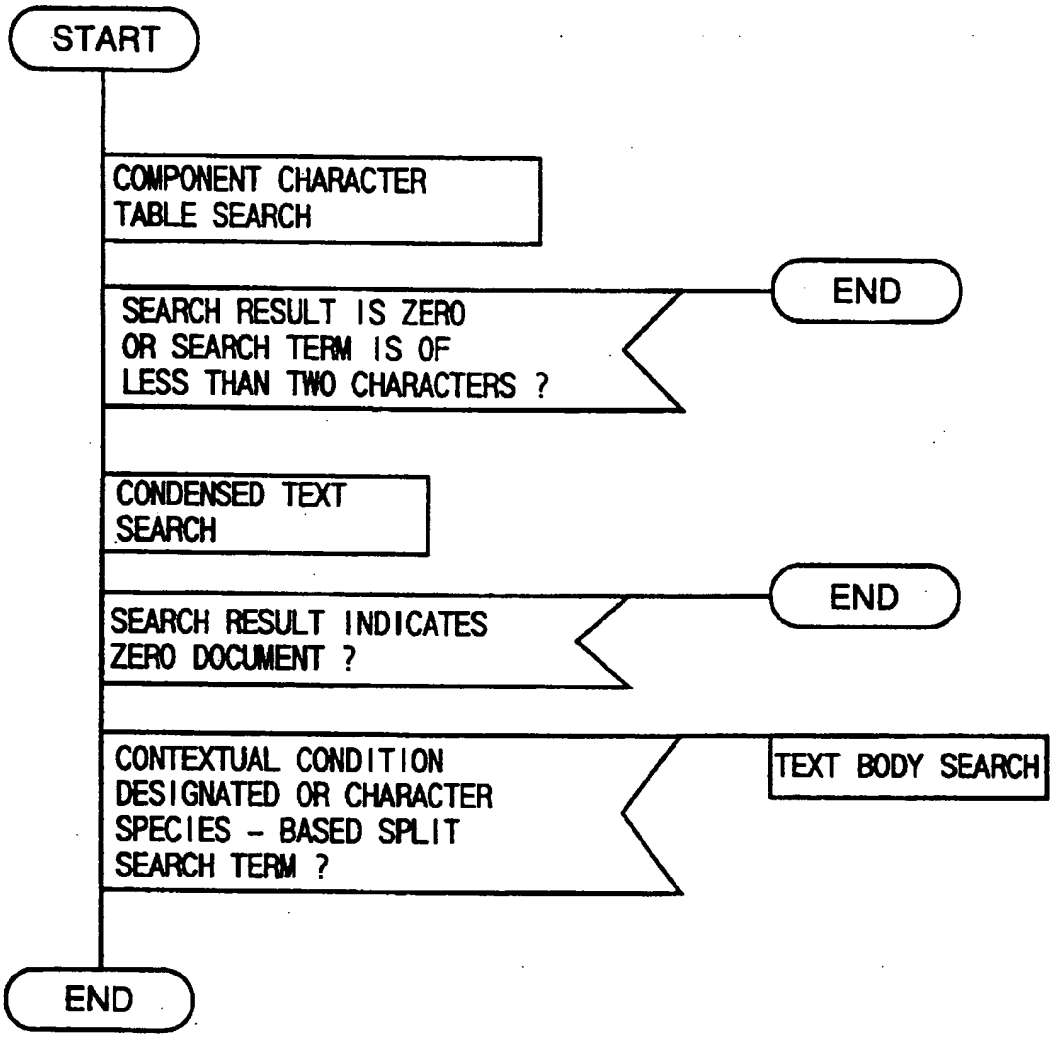

START
COMPONENT CHARACTER TABLE SEARCH
SEARCH RESULT IS ZERO OR SEARCH TERM IS OF LESS THAN TWO CHARACTERS ?
END
CONDENSED TEXT SEARCH
SEARCH RESULT INDICATES ZERO DOCUMENT ?
END
CONTEXTUAL CONDITION DESIGNATED OR CHARACTER SPECIES - BASED SPLIT SEARCH TERM ?
TEXT BODY SEARCH
END

FIG. 42

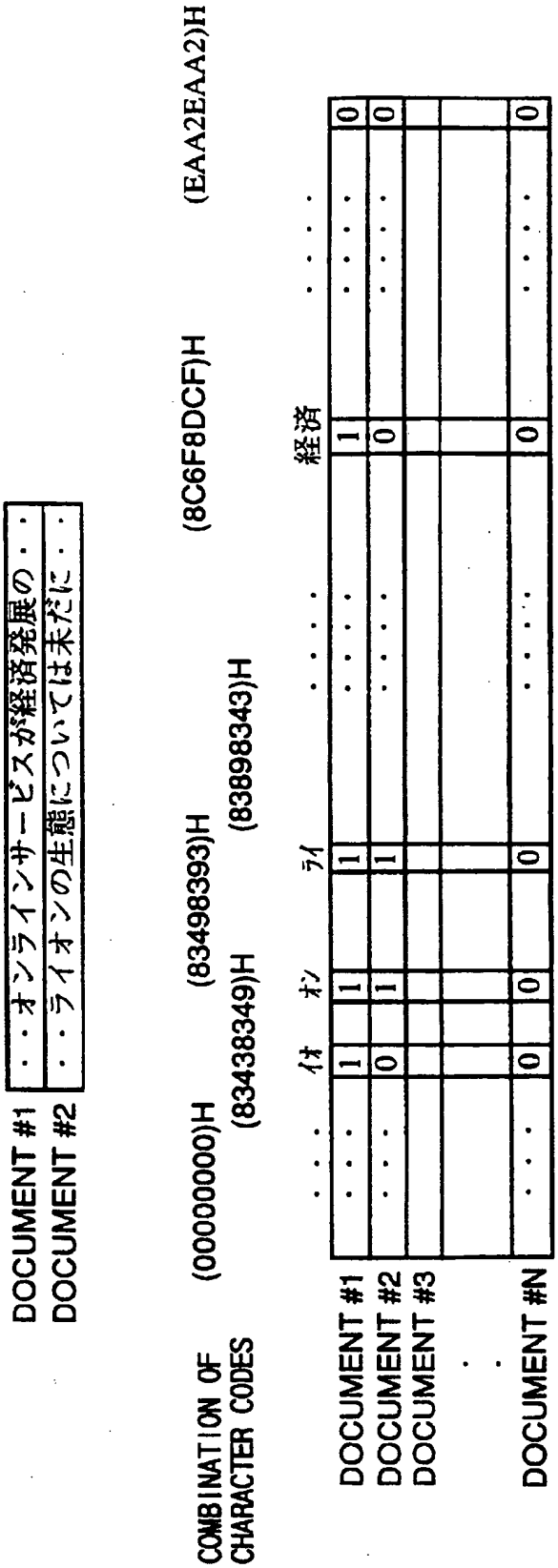

DOCUMENT #1 ・・オンラインサービスが経済発展の・・
DOCUMENT #2 ・・ライオンの生態については未だに・・
COMBINATION OF CHARACTER CODES
(00000000)H
(83438349)H
(83498393)H
(83898343)H
(8C6F8DCF)H
(EAA2EAA2)H
・・・・ 体 オン ライ ・・・・・ 経済 ・・・・・
DOCUMENT #1 ・・・ 1 1 1 ・・・・・ 1 ・・・・・ 0
DOCUMENT #2 ・・・ 0 1 1 ・・・・・ 0 ・・・・・ 0
DOCUMENT #3
DOCUMENT #N ・・・・ 0 0 0 ・・・・・ 0 ・・・・・ 0

FIG. 43

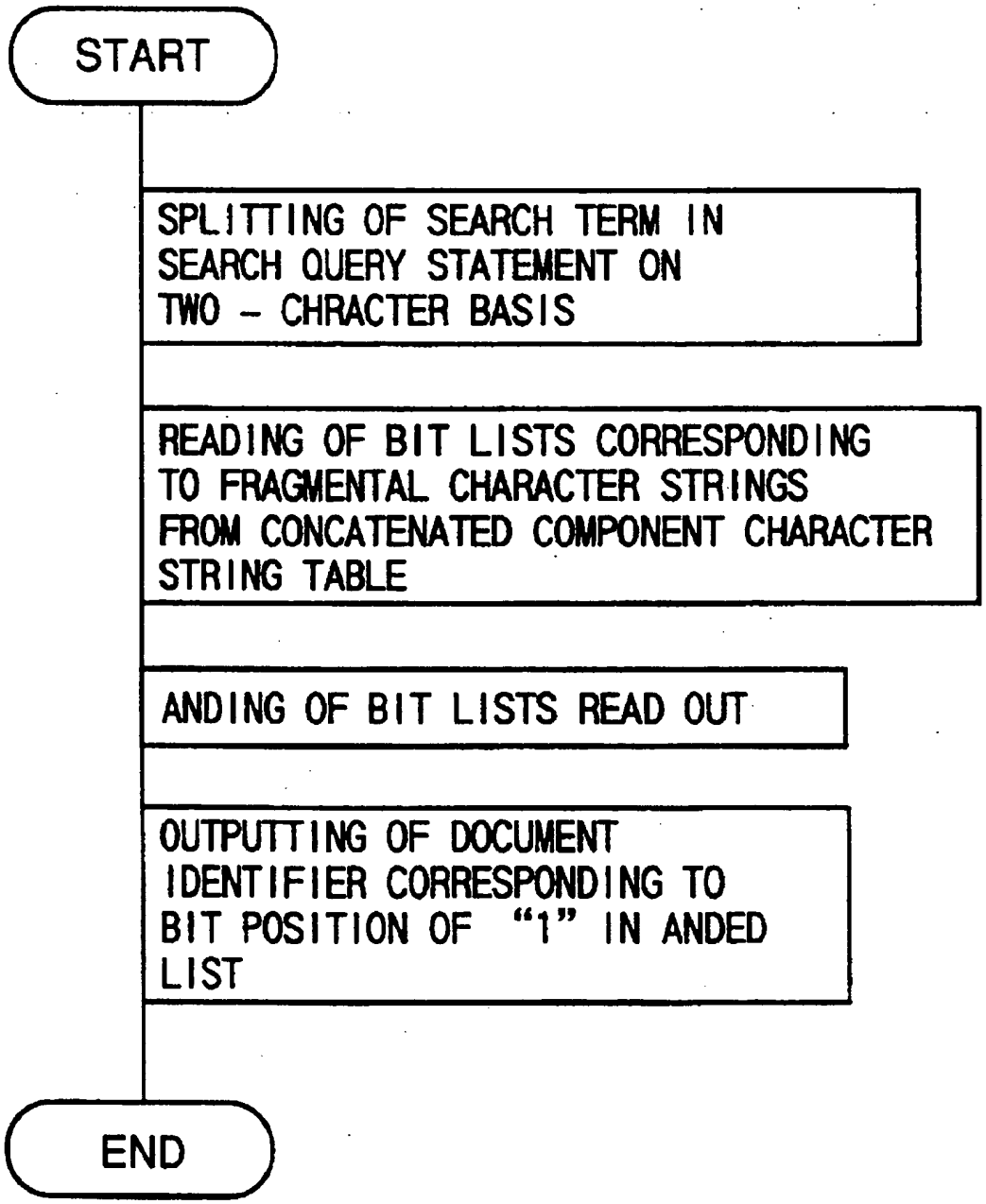

START
SPLITTING OF SEARCH TERM IN SEARCH QUERY STATEMENT ON TWO - CHRACTER BASIS
READING OF BIT LISTS CORRESPONDING TO FRAGMENTAL CHARACTER STRINGS FROM CONCATENATED COMPONENT CHARACTER STRING TABLE
ANDING OF BIT LISTS READ OUT
OUTPUTTING OF DOCUMENT IDENTIFIER CORRESPONDING TO BIT POSITION OF "1" IN ANDED LIST
END

FIG. 44

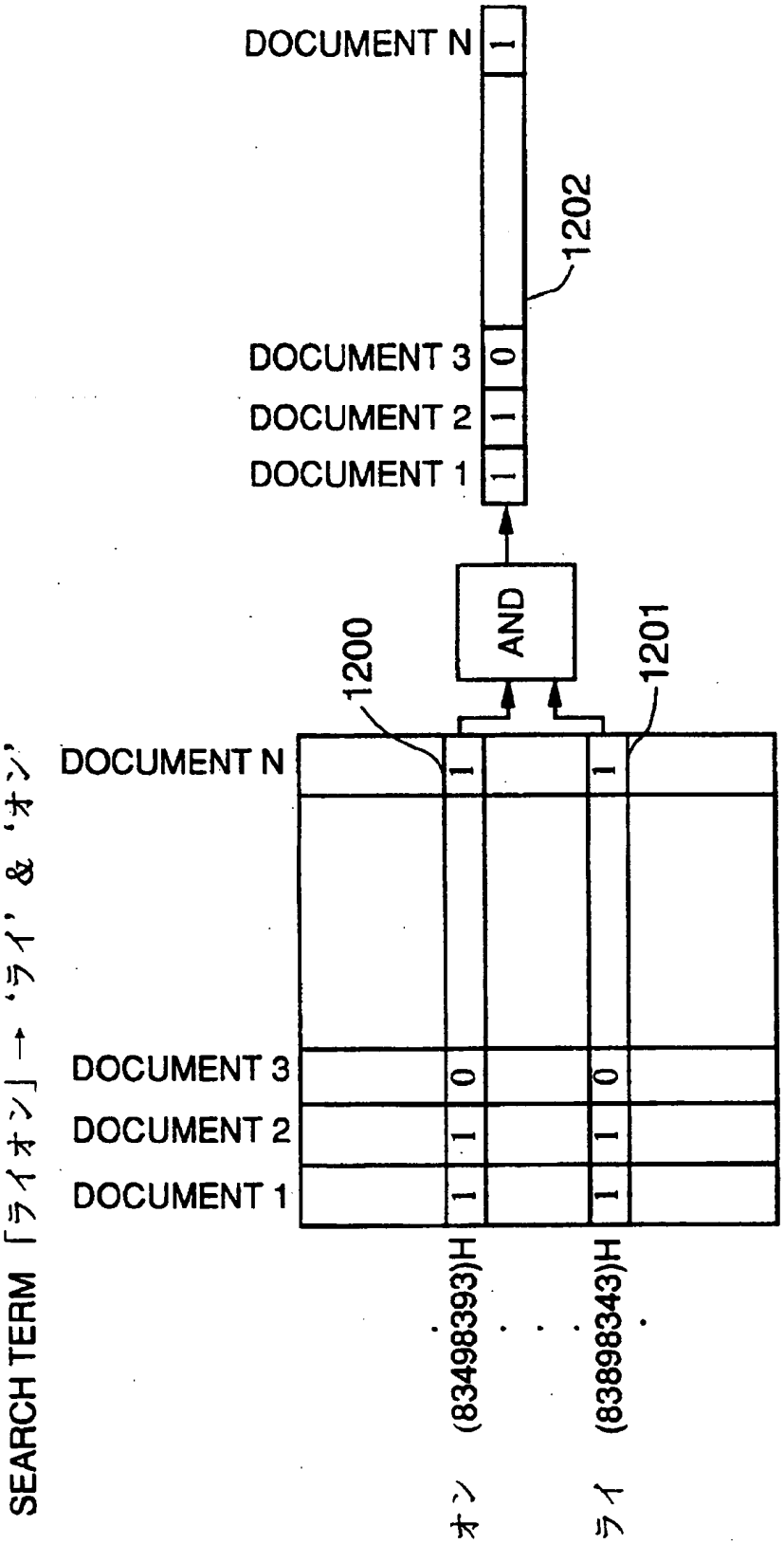

DOCUMENT N
1
1202
DOCUMENT 3
0
DOCUMENT 2
1
DOCUMENT 1
1
AND
1200
1201
DOCUMENT N
1
1
DOCUMENT 3
0
0
DOCUMENT 2
1
1
DOCUMENT 1
1
1
SEARCH TERM「ライオン」→「ライ」&「オン」
(8349839 3)H
(8389834 3)H
オン
ライ

FIG. 45

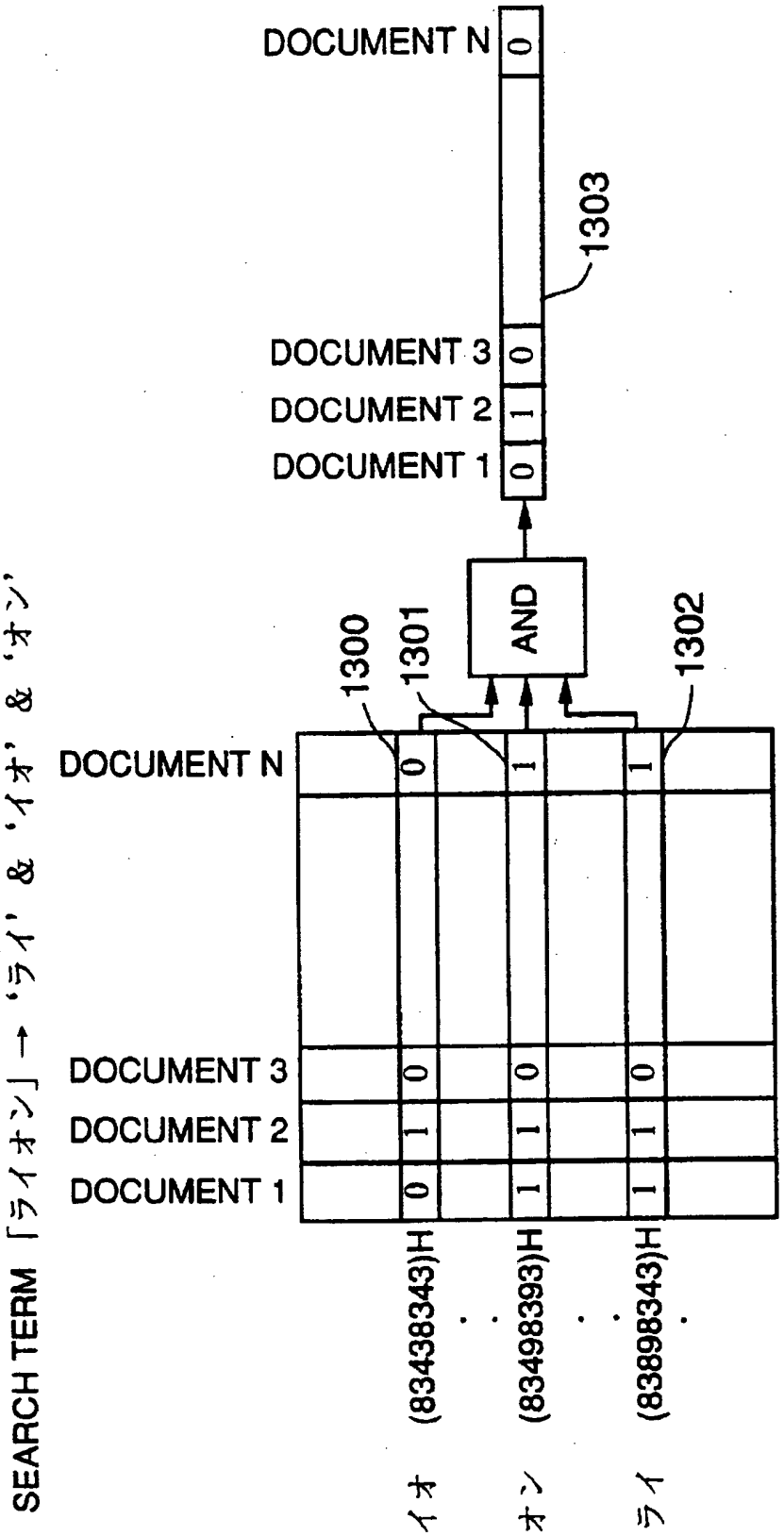

SEARCH TERM「ライオン」→'ライ' & 'イオ' & 'オン'
DOCUMENT N
DOCUMENT 3
DOCUMENT 2
DOCUMENT 1
1300
1301
1302
AND
1303
(83438343)H
(83498393)H
(83898343)H
イオ
オン
ライ

FIG. 46

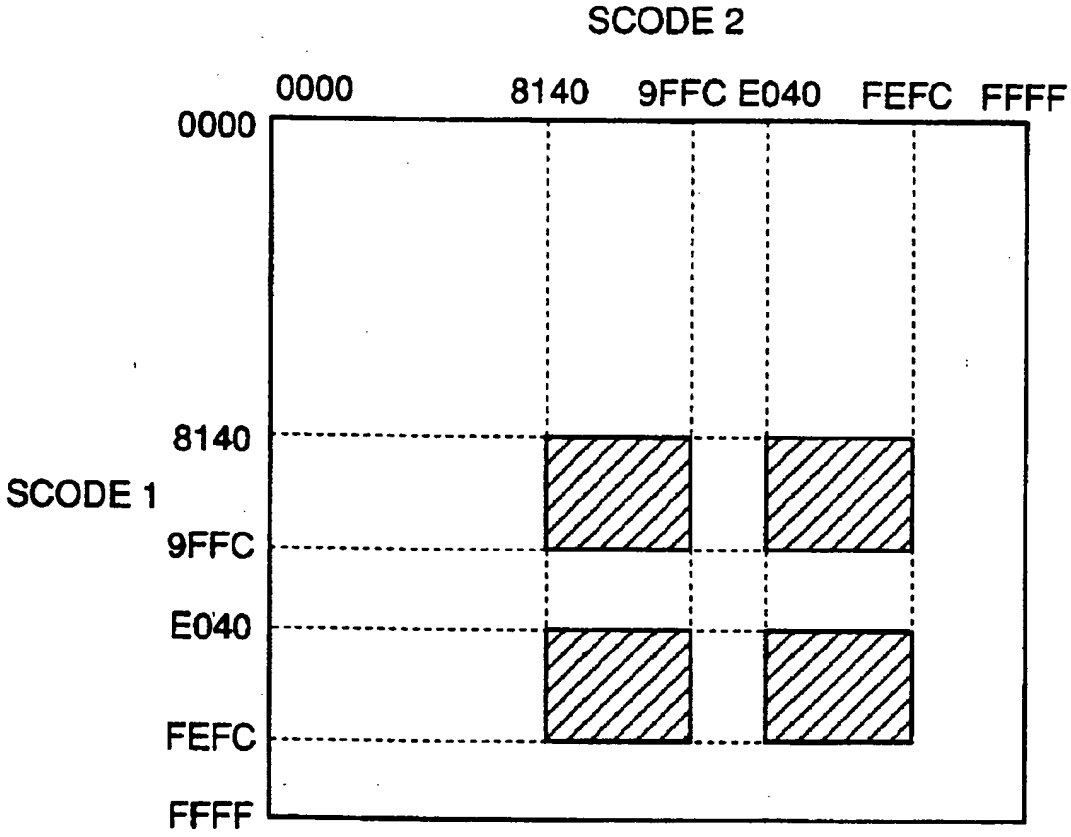

SCODE 2
0000
8140
9FFC
E040
FEFC
FFFF
SCODE 1
0000
8140
9FFC
E040
FEFC
FFFF

FIG. 47

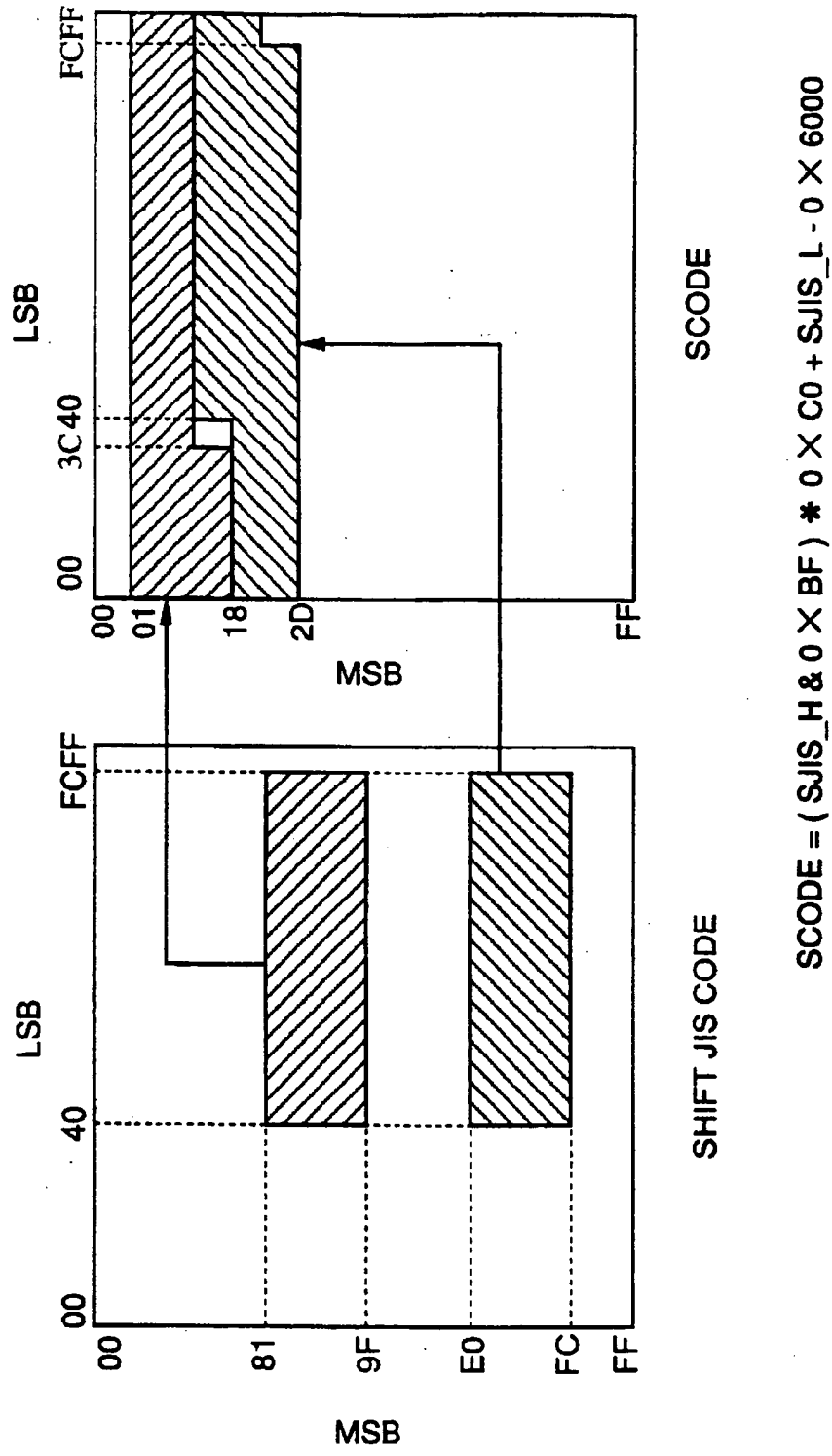

FCFF
3C 40
00
LSB
00
01
18
2D
FF
MSB
SCODE
FCFF
40
00
LSB
00
81
9F
E0
FC
FF
MSB
SHIFT JIS CODE
SCODE = ( SJIS_H & 0 X BF ) * 0 X C0 + SJIS_L - 0 X 6000

FIG. 48

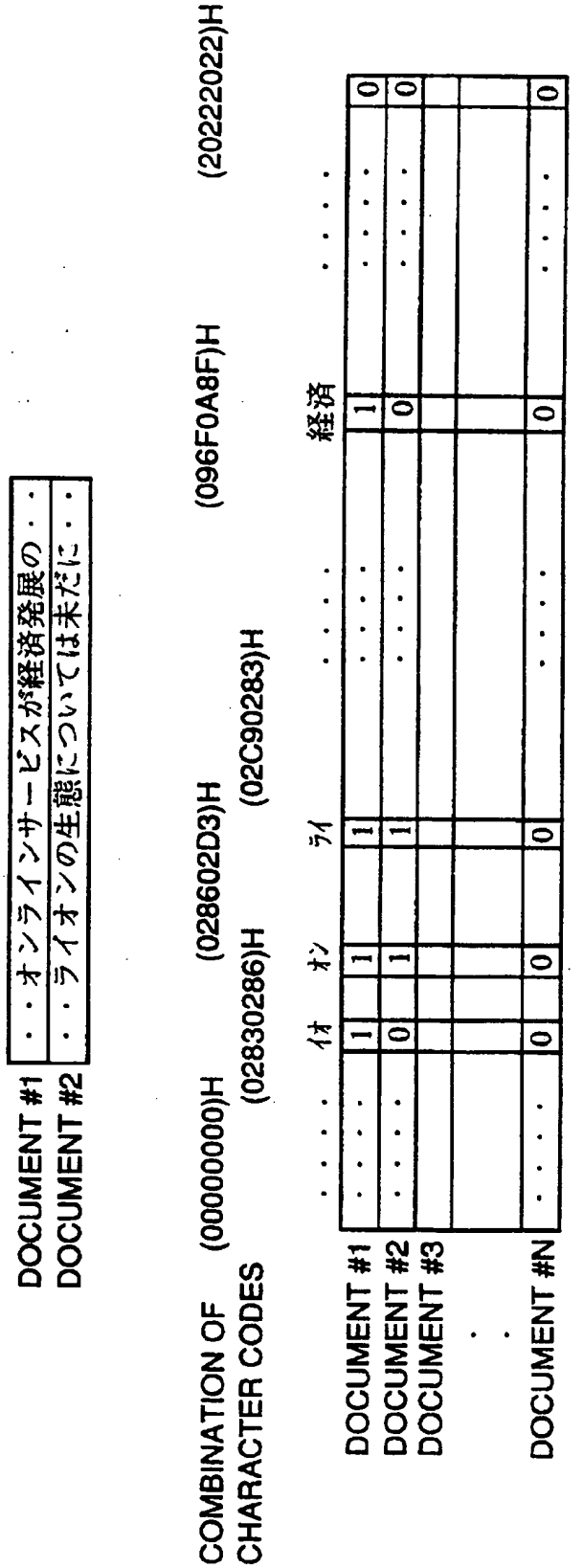

DOCUMENT #1
DOCUMENT #2
・・オンラインサービスが経済発展の・・
・・ライオンの生態については未だに・・
COMBINATION OF CHARACTER CODES
(00000000)H
(02830286)H
(028602D3)H
(02C90283)H
(096F0A8F)H
(20222022)H
・・・・・ 仕 オン ライ ・・・・・ 経済 ・・・・・
DOCUMENT #1
DOCUMENT #2
DOCUMENT #3
DOCUMENT #N

FIG. 49

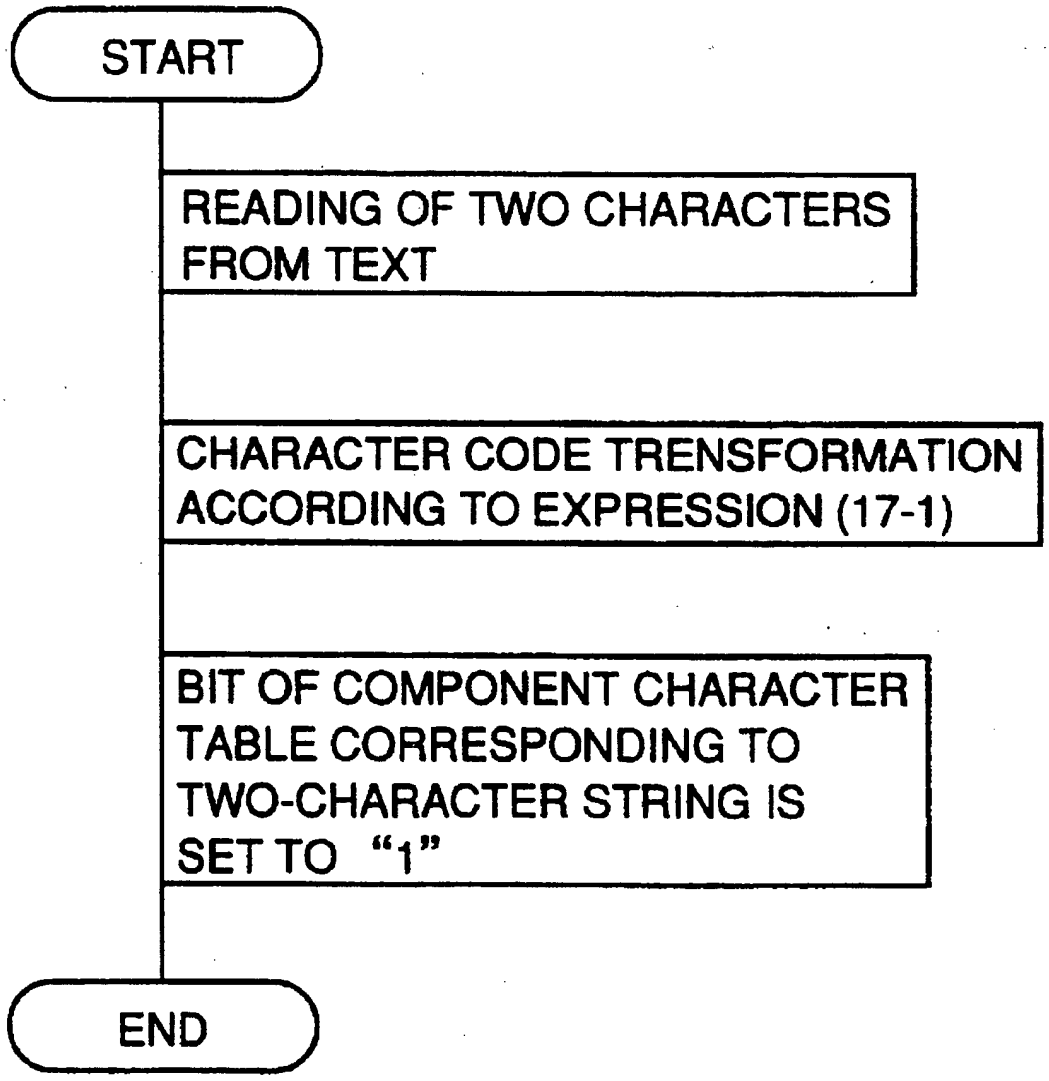

START
READING OF TWO CHARACTERS FROM TEXT
CHARACTER CODE TRENSFORMATION ACCORDING TO EXPRESSION (17-1)
BIT OF COMPONENT CHARACTER TABLE CORRESPONDING TO TWO-CHARACTER STRING IS SET TO "1"
END

FIG. 50

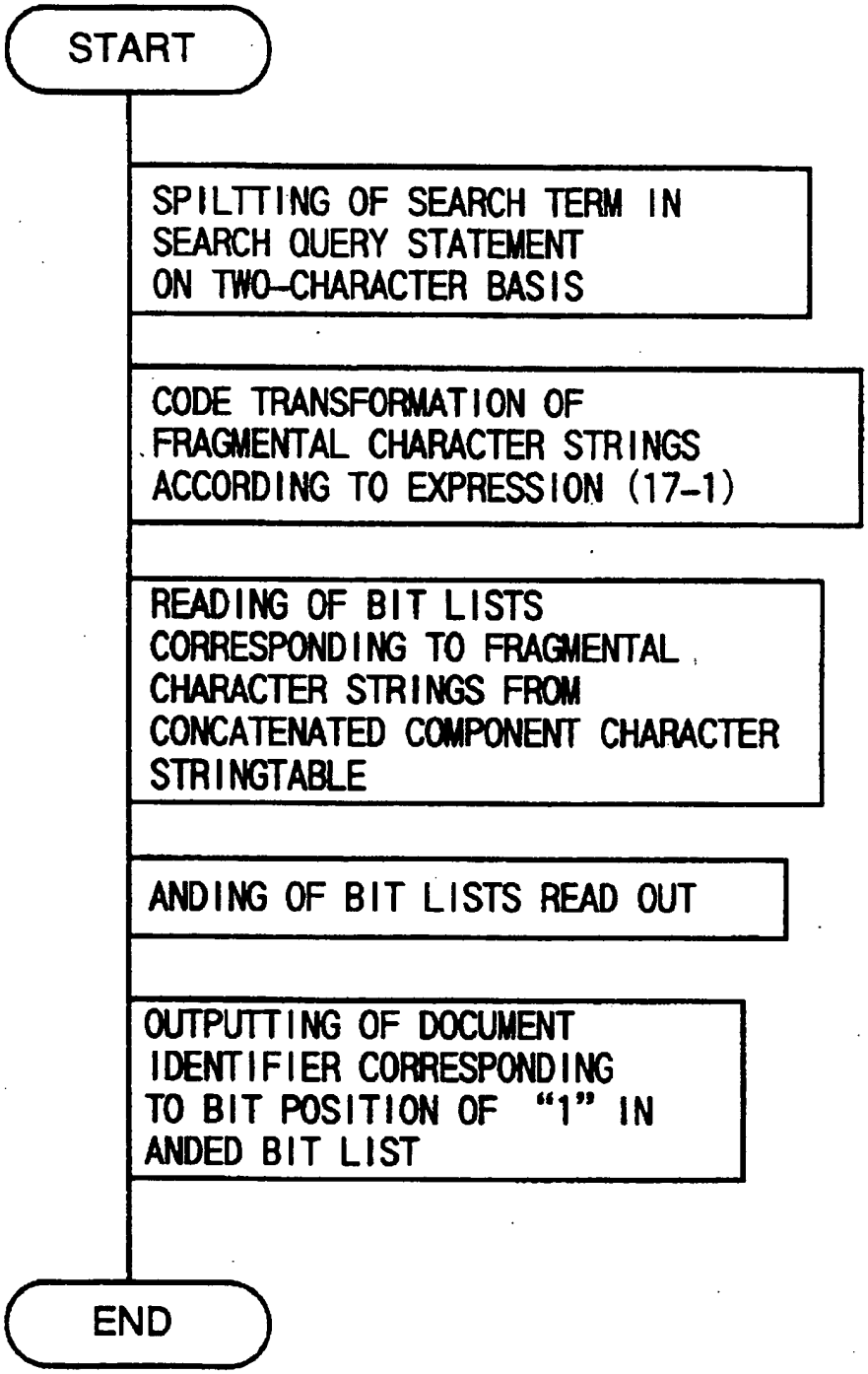

START
SPILTTING OF SEARCH TERM IN SEARCH QUERY STATEMENT ON TWO-CHARACTER BASIS
CODE TRANSFORMATION OF FRAGMENTAL CHARACTER STRINGS ACCORDING TO EXPRESSION (17-1)
READING OF BIT LISTS CORRESPONDING TO FRAGMENTAL CHARACTER STRINGS FROM CONCATENATED COMPONENT CHARACTER STRINGTABLE
ANDING OF BIT LISTS READ OUT
OUTPUTTING OF DOCUMENT IDENTIFIER CORRESPONDING TO BIT POSITION OF "1" IN ANDED BIT LIST
END

FIG. 51

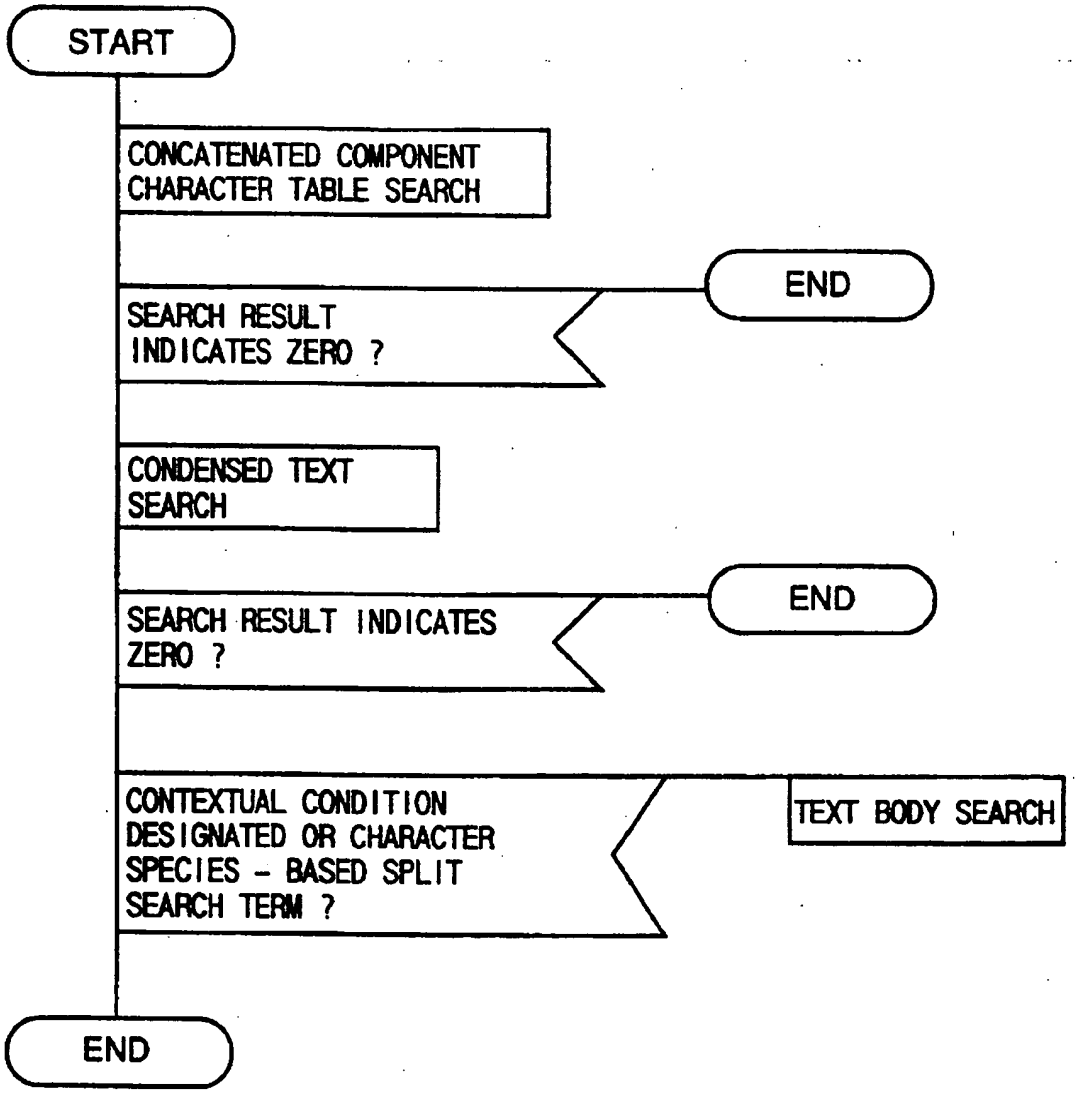

START
CONCATENATED COMPONENT CHARACTER TABLE SEARCH
END
SEARCH RESULT INDICATES ZERO ?
CONDENSED TEXT SEARCH
END
SEARCH RESULT INDICATES ZERO ?
CONTEXTUAL CONDITION DESIGNATED OR CHARACTER SPECIES - BASED SPLIT SEARCH TERM ?
TEXT BODY SEARCH
END

FIG. 52

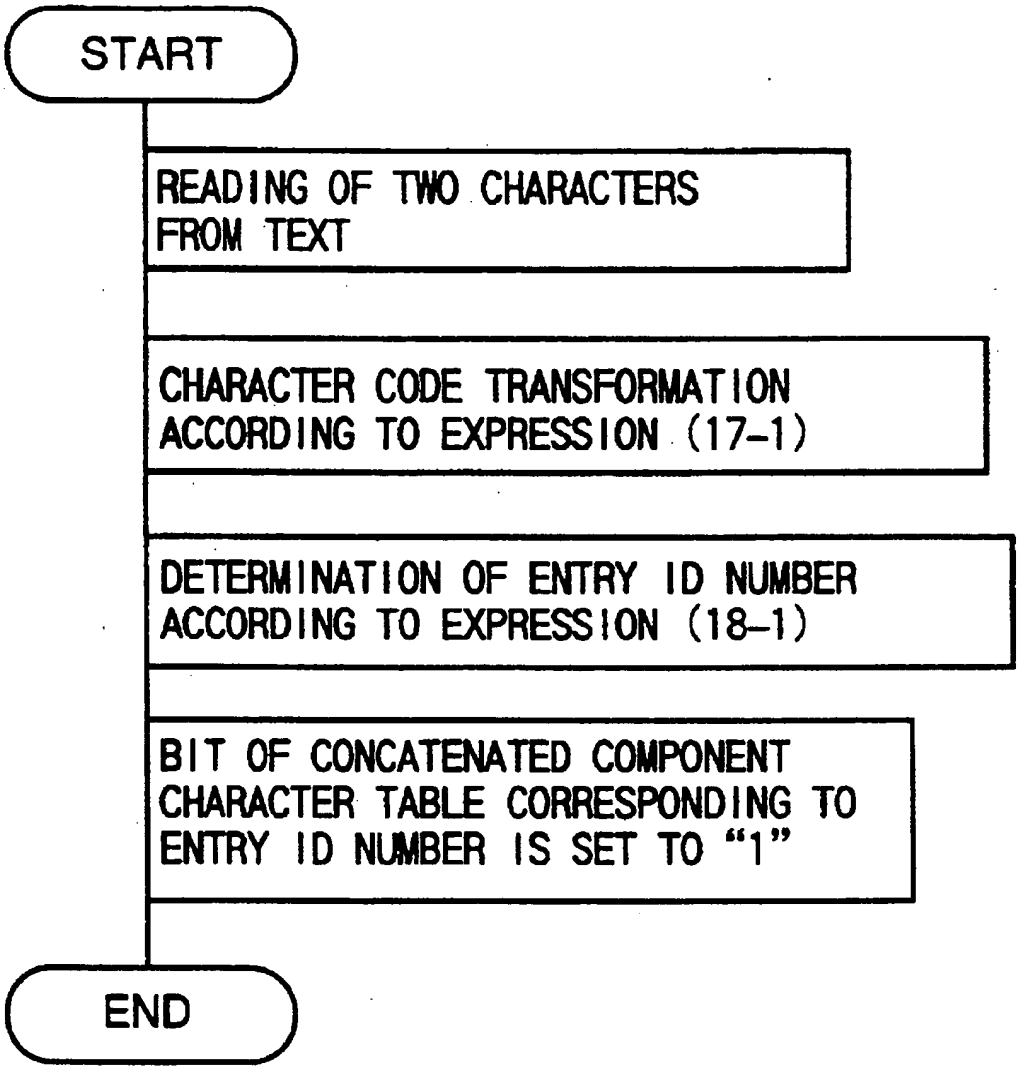

START
READING OF TWO CHARACTERS FROM TEXT
CHARACTER CODE TRANSFORMATION ACCORDING TO EXPRESSION (17-1)
DETERMINATION OF ENTRY ID NUMBER ACCORDING TO EXPRESSION (18-1)
BIT OF CONCATENATED COMPONENT CHARACTER TABLE CORRESPONDING TO ENTRY ID NUMBER IS SET TO "1"
END

FIG. 53

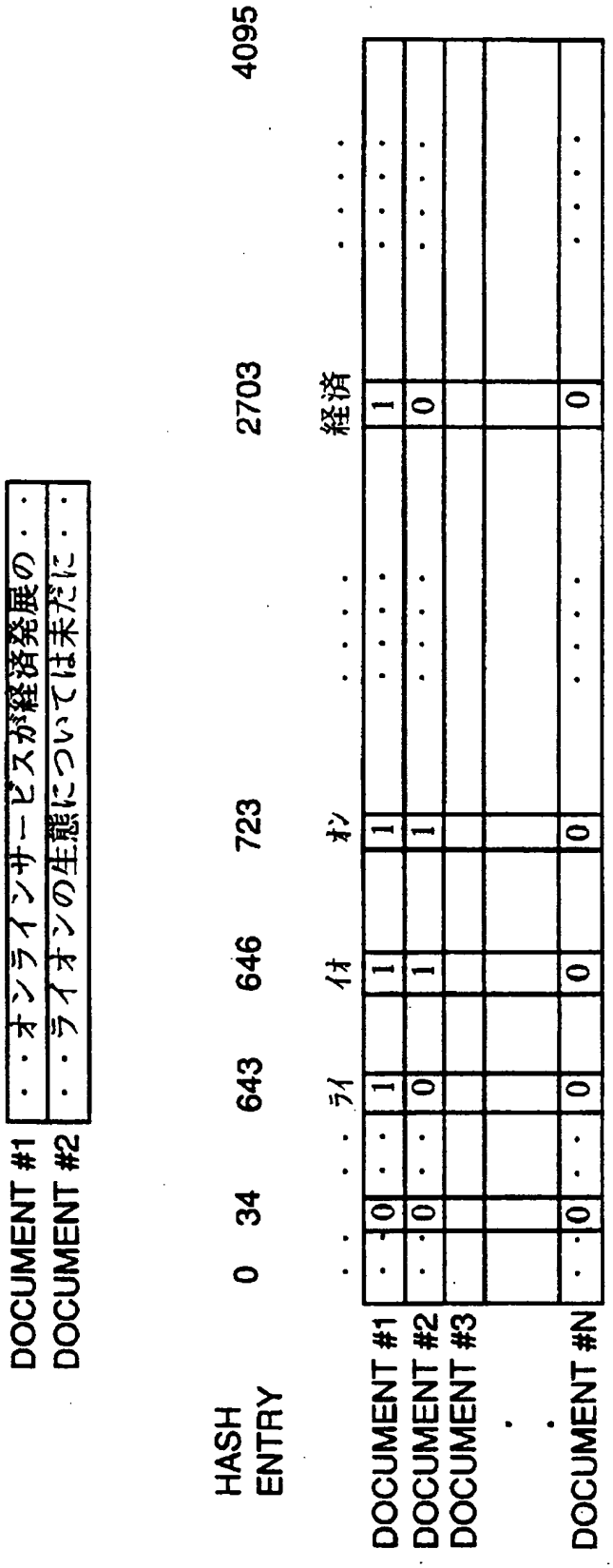

DOCUMENT #1
DOCUMENT #2
・・オンラインサービスが経済発展の・
・・ライオンの生態については未だに・
HASH ENTRY
0
34
643
646
723
2703
4095
経済
DOCUMENT #1
DOCUMENT #2
DOCUMENT #3
DOCUMENT #N

FIG. 54

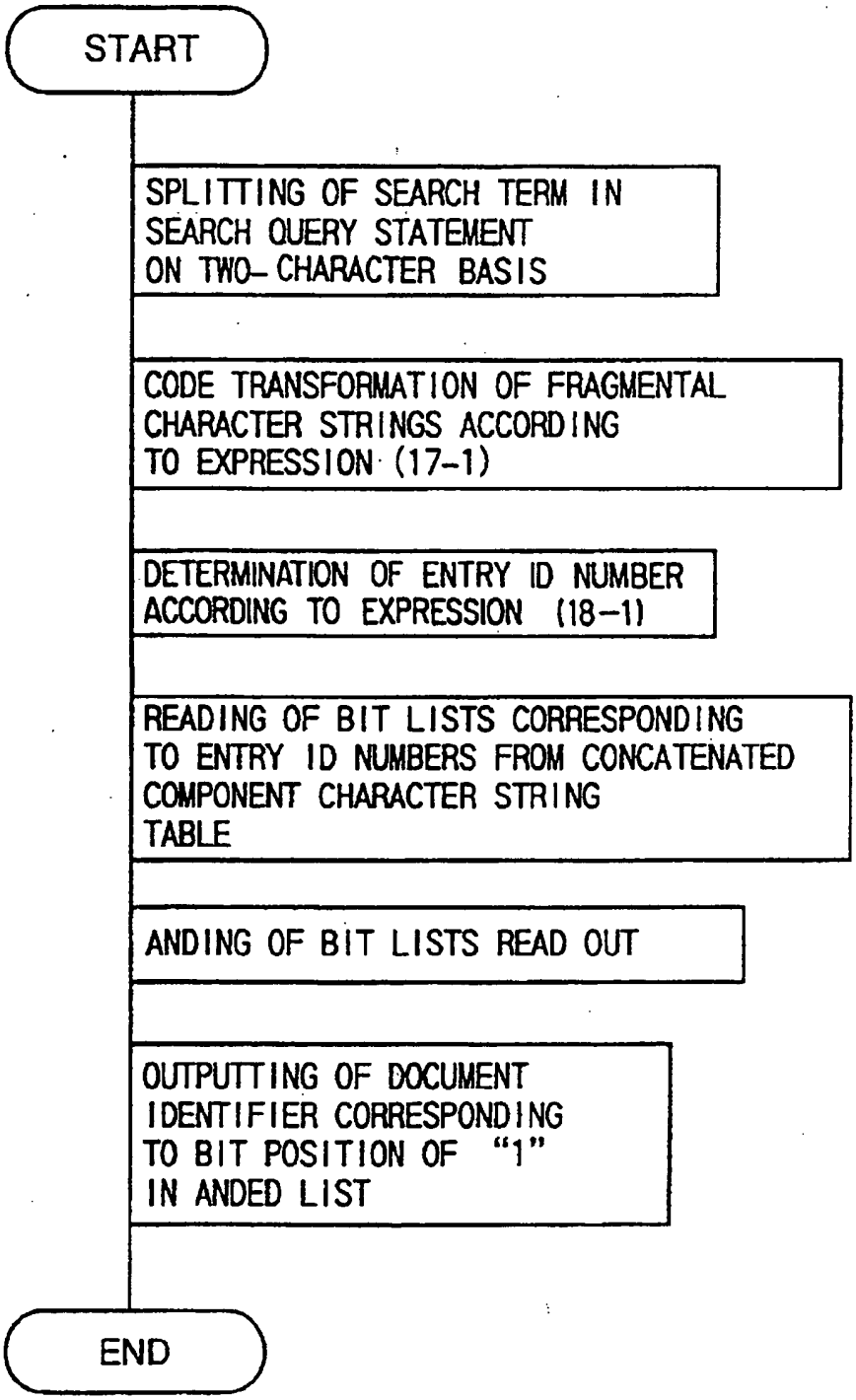

START
SPLITTING OF SEARCH TERM IN SEARCH QUERY STATEMENT ON TWO-CHARACTER BASIS
CODE TRANSFORMATION OF FRAGMENTAL CHARACTER STRINGS ACCORDING TO EXPRESSION (17-1)
DETERMINATION OF ENTRY ID NUMBER ACCORDING TO EXPRESSION (18-1)
READING OF BIT LISTS CORRESPONDING TO ENTRY ID NUMBERS FROM CONCATENATED COMPONENT CHARACTER STRING TABLE
ANDING OF BIT LISTS READ OUT
OUTPUTTING OF DOCUMENT IDENTIFIER CORRESPONDING TO BIT POSITION OF "1" IN ANDED LIST
END

FIG. 55

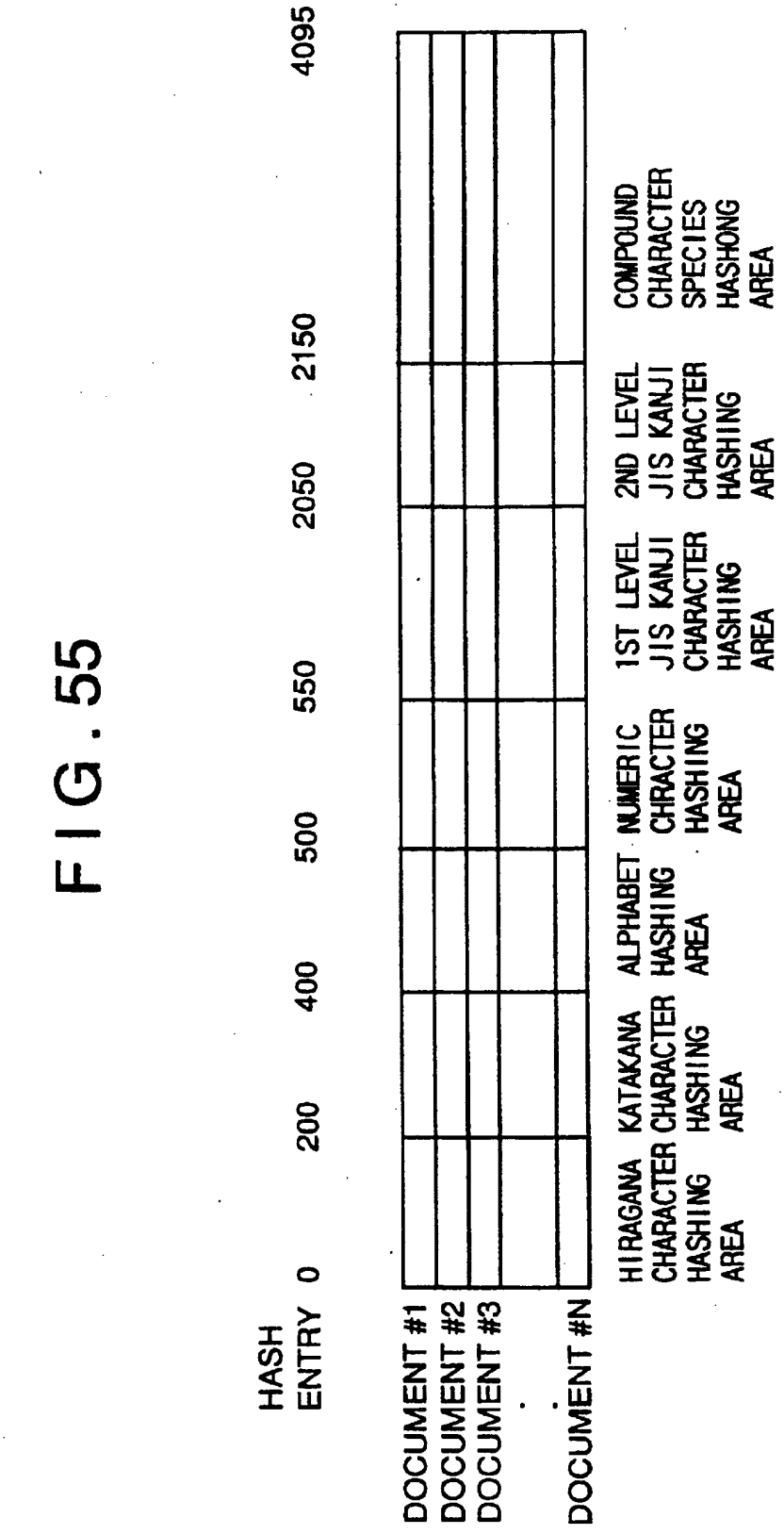

HASH ENTRY
0
200
400
500
550
2050
2150
4095
DOCUMENT #1
DOCUMENT #2
DOCUMENT #3
DOCUMENT #N
HIRAGANA CHARACTER HASHING AREA
KATAKANA CHARACTER HASHING AREA
ALPHABET HASHING AREA
NUMERIC CHRACTER HASHING AREA
1ST LEVEL JIS KANJI CHARACTER HASHING AREA
2ND LEVEL JIS KANJI CHARACTER HASHING AREA
COMPOUND CHARACTER SPECIES HASHONG AREA

FIG. 56

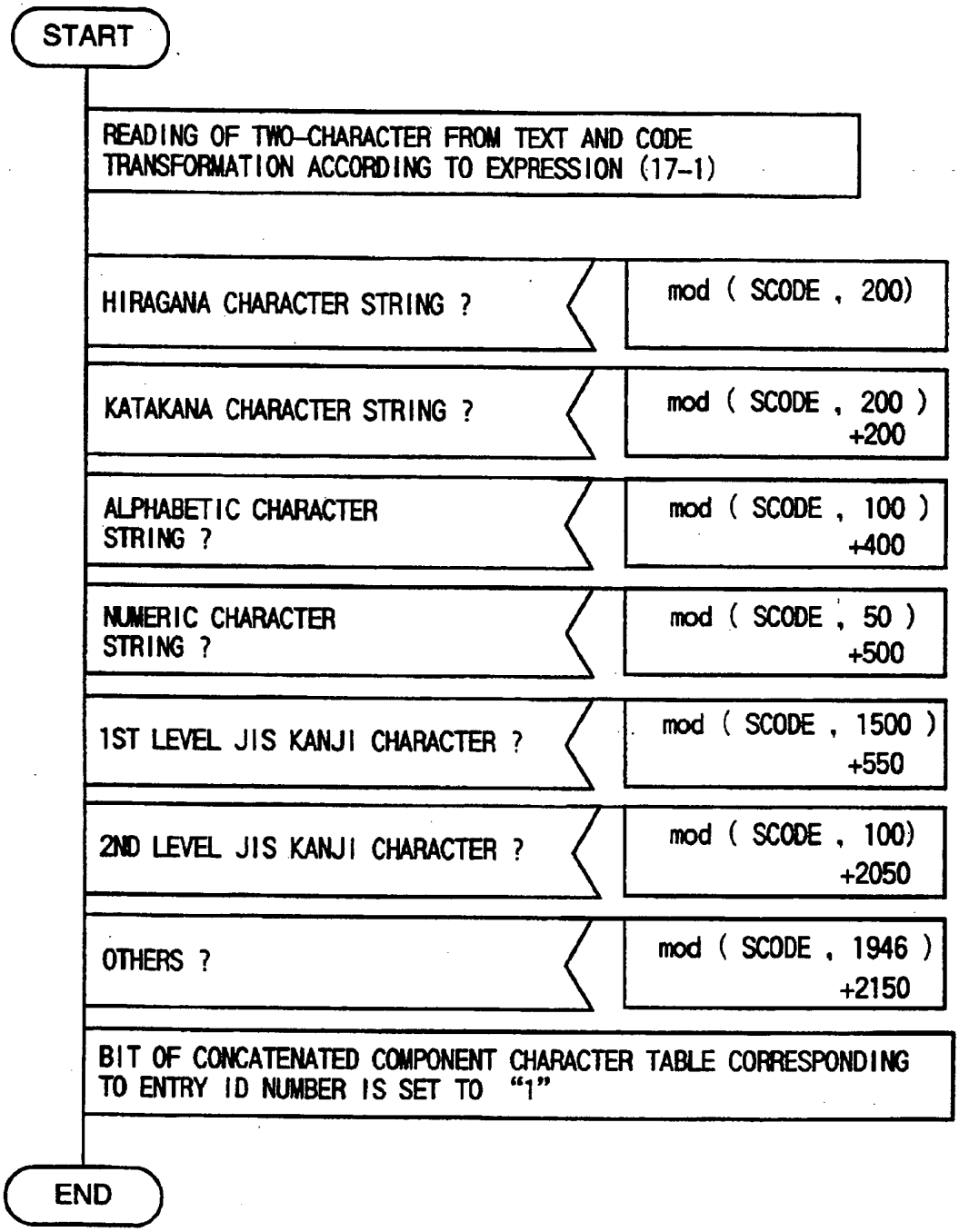

START
READING OF TWO-CHARACTER FROM TEXT AND CODE TRANSFORMATION ACCORDING TO EXPRESSION (17-1)
HIRAGANA CHARACTER STRING ?
mod ( SCODE , 200)
KATAKANA CHARACTER STRING ?
mod ( SCODE , 200 ) +200
ALPHABETIC CHARACTER STRING ?
mod ( SCODE , 100 ) +400
NUMERIC CHARACTER STRING ?
mod ( SCODE , 50 ) +500
1ST LEVEL JIS KANJI CHARACTER ?
mod ( SCODE , 1500 ) +550
2ND LEVEL JIS KANJI CHARACTER ?
mod ( SCODE , 100) +2050
OTHERS ?
mod ( SCODE , 1946 ) +2150
BIT OF CONCATENATED COMPONENT CHARACTER TABLE CORRESPONDING TO ENTRY ID NUMBER IS SET TO "1"
END

FIG. 57

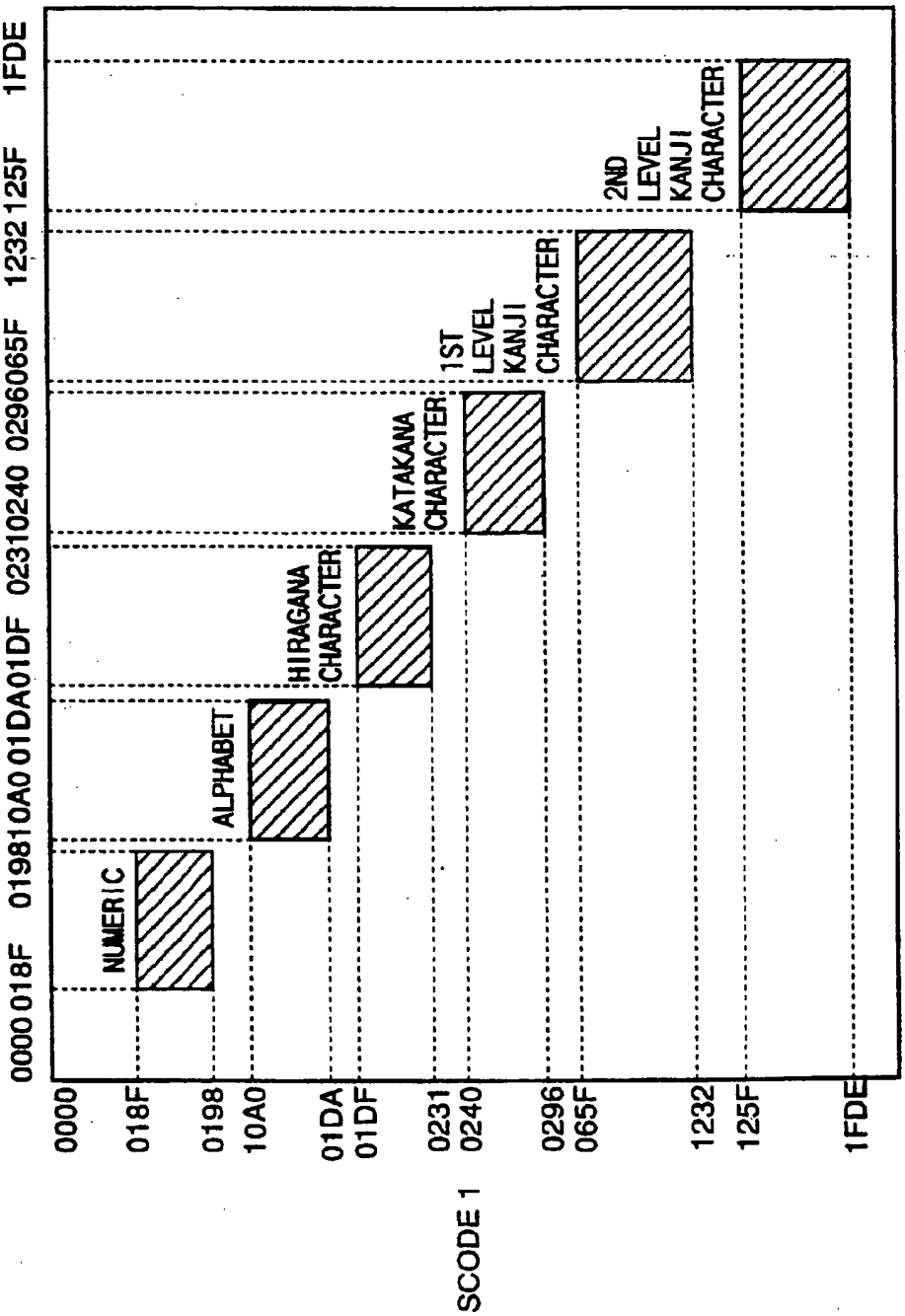

SCODE 2
0000 018F 0198 10A0 01DA 01DF 0231 0240 0296 065F 1232 125F 1FDE
NUMERIC
ALPHABET
HIRAGANA CHARACTER
KATAKANA CHARACTER
1ST LEVEL KANJI CHARACTER
2ND LEVEL KANJI CHARACTER
SCODE 1
0000 018F 0198 10A0 01DA 01DF 0231 0240 0296 065F 1232 125F 1FDE

FIG. 58

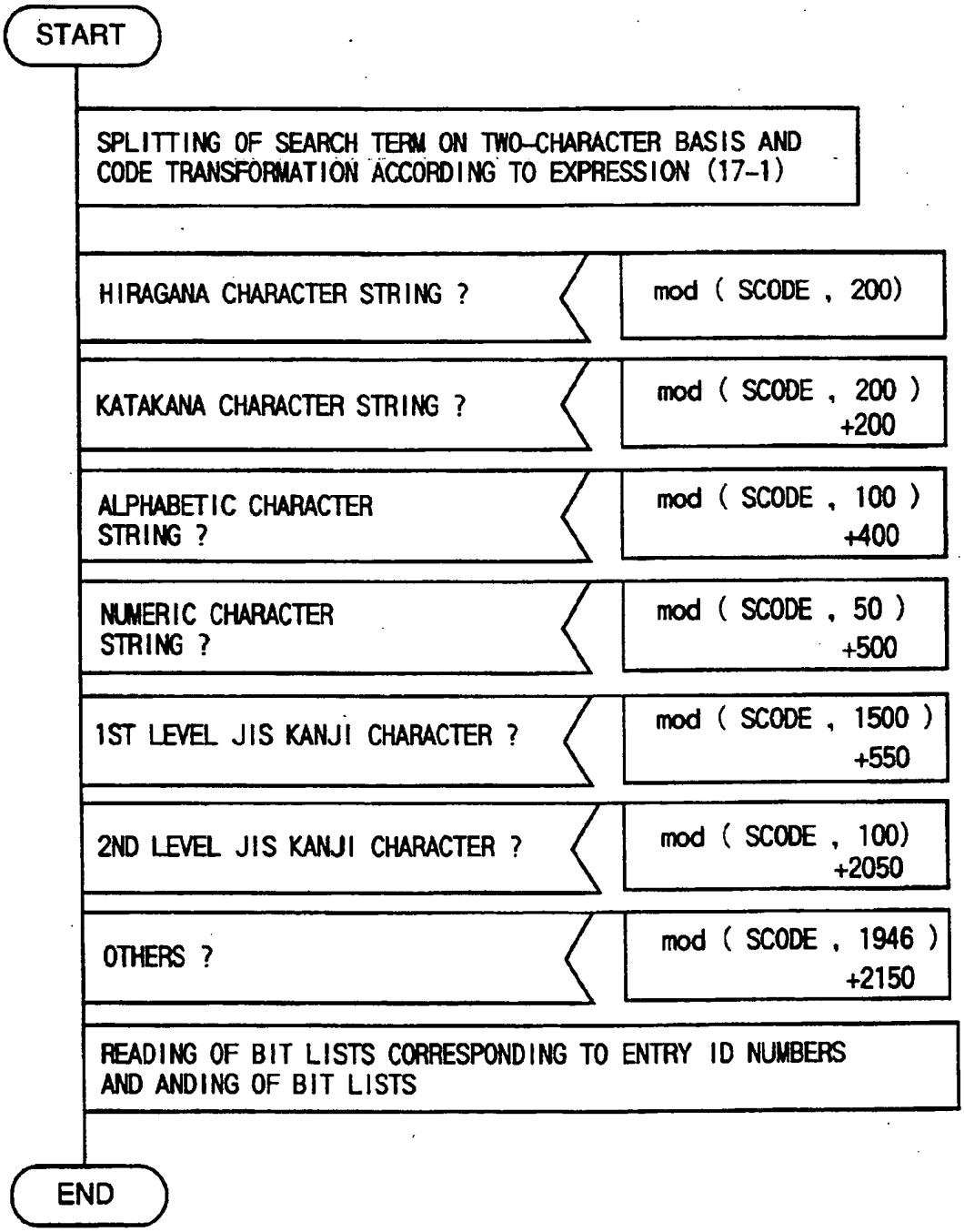

START
SPLITTING OF SEARCH TERM ON TWO-CHARACTER BASIS AND CODE TRANSFORMATION ACCORDING TO EXPRESSION (17-1)
HIRAGANA CHARACTER STRING ?
mod ( SCODE , 200)
KATAKANA CHARACTER STRING ?
mod ( SCODE , 200 ) +200
ALPHABETIC CHARACTER STRING ?
mod ( SCODE , 100 ) +400
NUMERIC CHARACTER STRING ?
mod ( SCODE , 50 ) +500
1ST LEVEL JIS KANJI CHARACTER ?
mod ( SCODE , 1500 ) +550
2ND LEVEL JIS KANJI CHARACTER ?
mod ( SCODE , 100) +2050
OTHERS ?
mod ( SCODE , 1946 ) +2150
READING OF BIT LISTS CORRESPONDING TO ENTRY ID NUMBERS AND ANDING OF BIT LISTS
END

FIG. 59

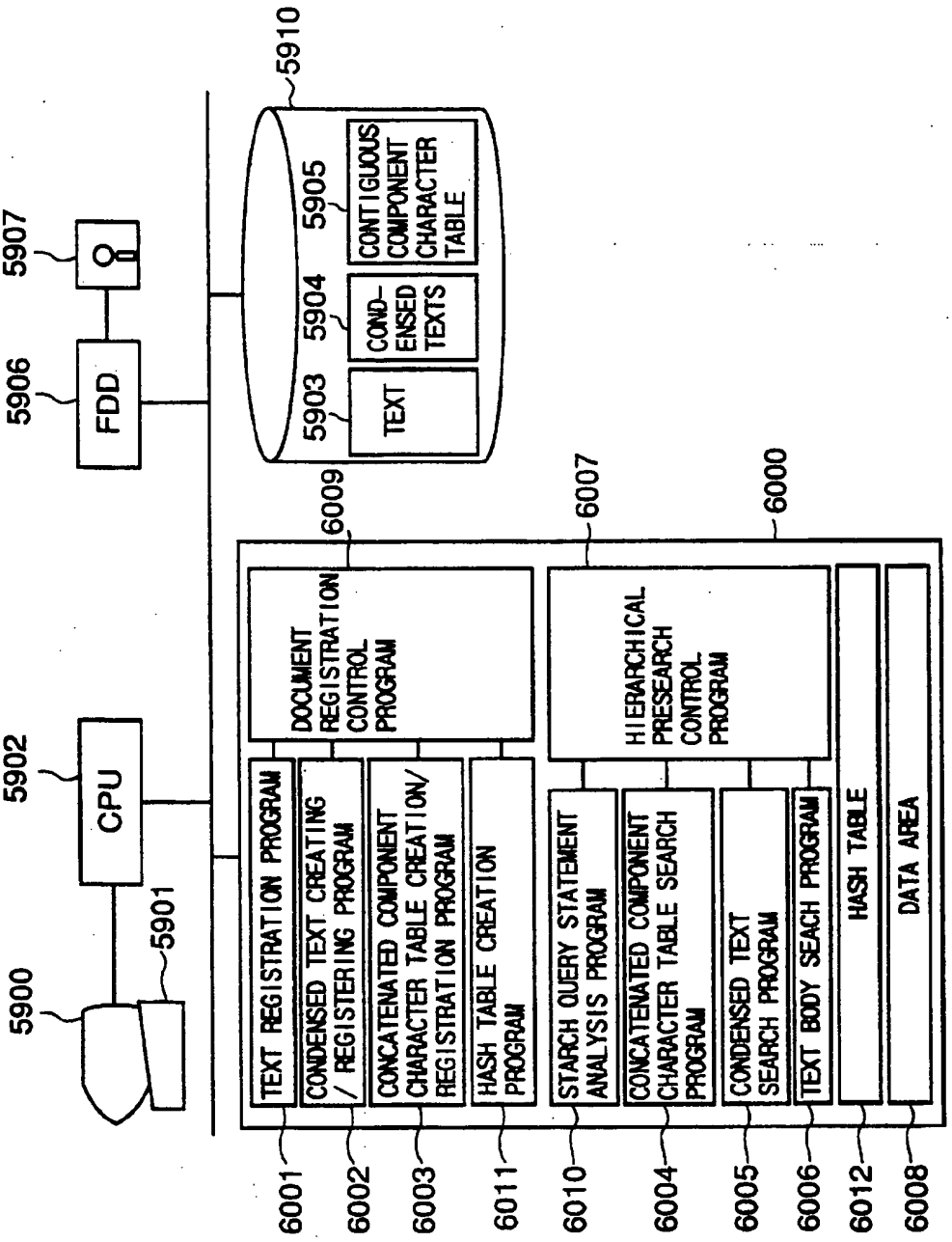

5910
5907
5906
FDD
5903
TEXT
5904
COND-ENSED TEXTS
5905
CONTIGUOUS COMPONENT CHARACTER TABLE
6009
6007
6000
5902
CPU
5901
5900
DOCUMENT REGISTRATION CONTROL PROGRAM
HIERARCHICAL PRESEARCH CONTROL PROGRAM
TEXT REGISTRATION PROGRAM
CONDENSED TEXT CREATING / REGISTERING PROGRAM
CONCATENATED COMPONENT CHARACTER TABLE CREATION/ REGISTRATION PROGRAM
HASH TABLE CREATION PROGRAM
STARCH QUERY STATEMENT ANALYSIS PROGRAM
CONCATENATED COMPONENT CHARACTER TABLE SEARCH PROGRAM
CONDENSED TEXT SEARCH PROGRAM
TEXT BODY SEACH PROGRAM
HASH TABLE
DATA AREA
6001
6002
6003
6011
6010
6004
6005
6006
6012
6008

FIG. 60

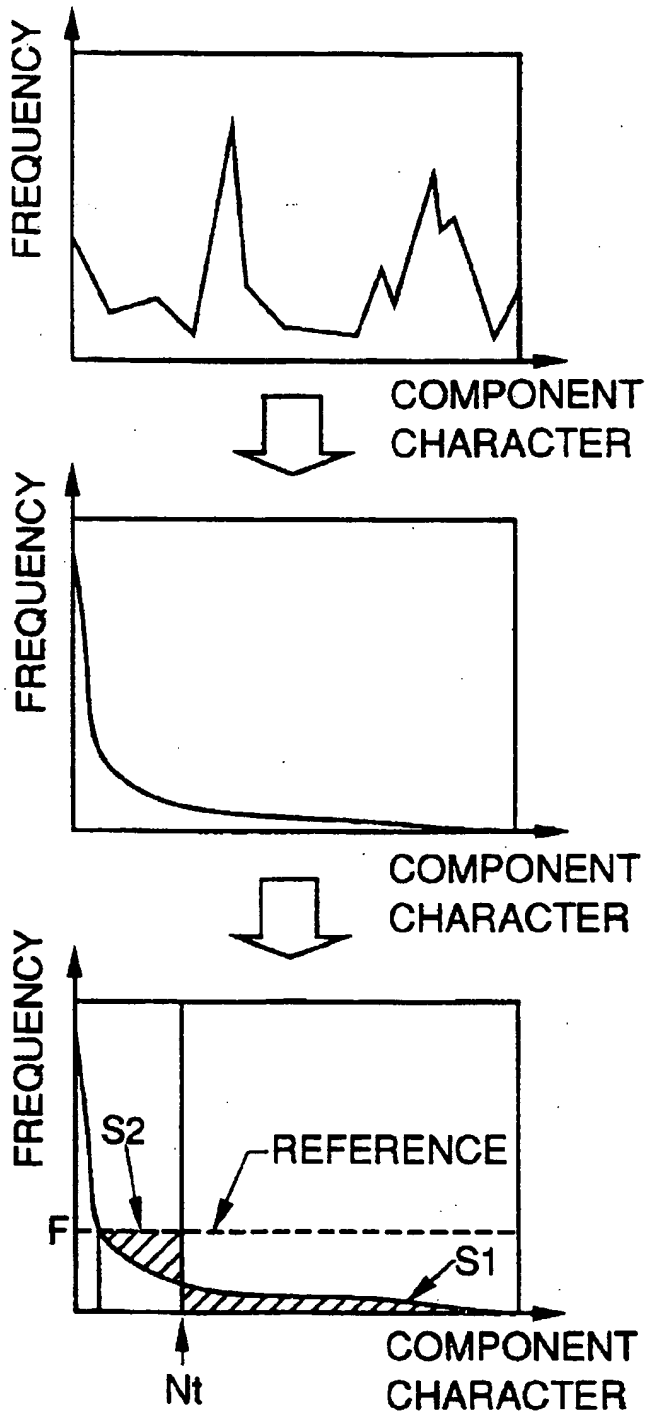

FREQUENCY
COMPONENT CHARACTER
FREQUENCY
COMPONENT CHARACTER
FREQUENCY
S2
REFERENCE
F
S1
Nt
COMPONENT CHARACTER

FIG. 61

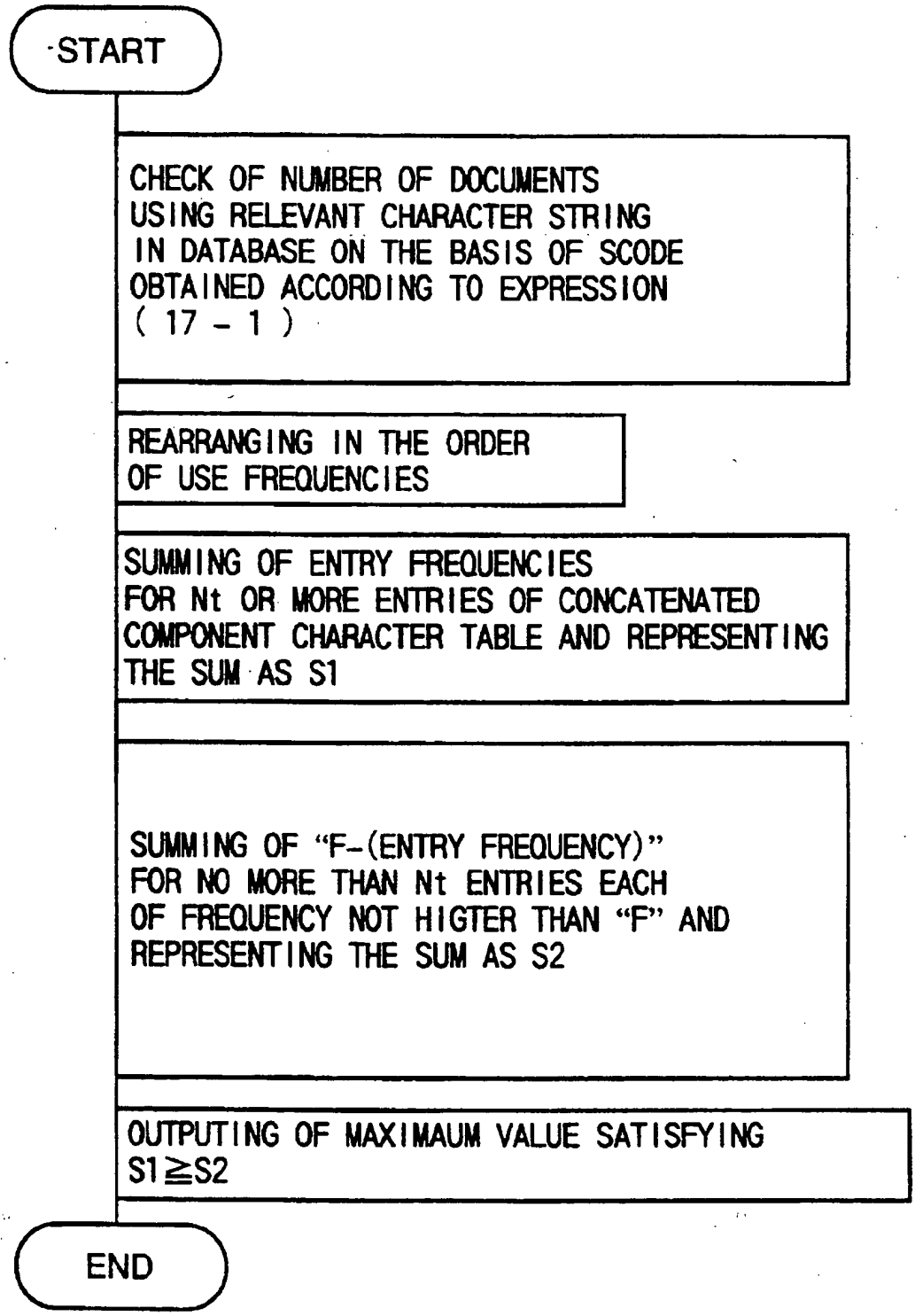

START
CHECK OF NUMBER OF DOCUMENTS USING RELEVANT CHARACTER STRING IN DATABASE ON THE BASIS OF SCODE OBTAINED ACCORDING TO EXPRESSION ( 17 - 1 )
REARRANGING IN THE ORDER OF USE FREQUENCIES
SUMMING OF ENTRY FREQUENCIES FOR Nt OR MORE ENTRIES OF CONCATENATED COMPONENT CHARACTER TABLE AND REPRESENTING THE SUM AS S1
SUMMING OF "F-(ENTRY FREQUENCY)" FOR NO MORE THAN Nt ENTRIES EACH OF FREQUENCY NOT HIGTER THAN "F" AND REPRESENTING THE SUM AS S2
OUTPUTING OF MAXIMAUM VALUE SATISFYING S1≧S2
END

FIG. 62

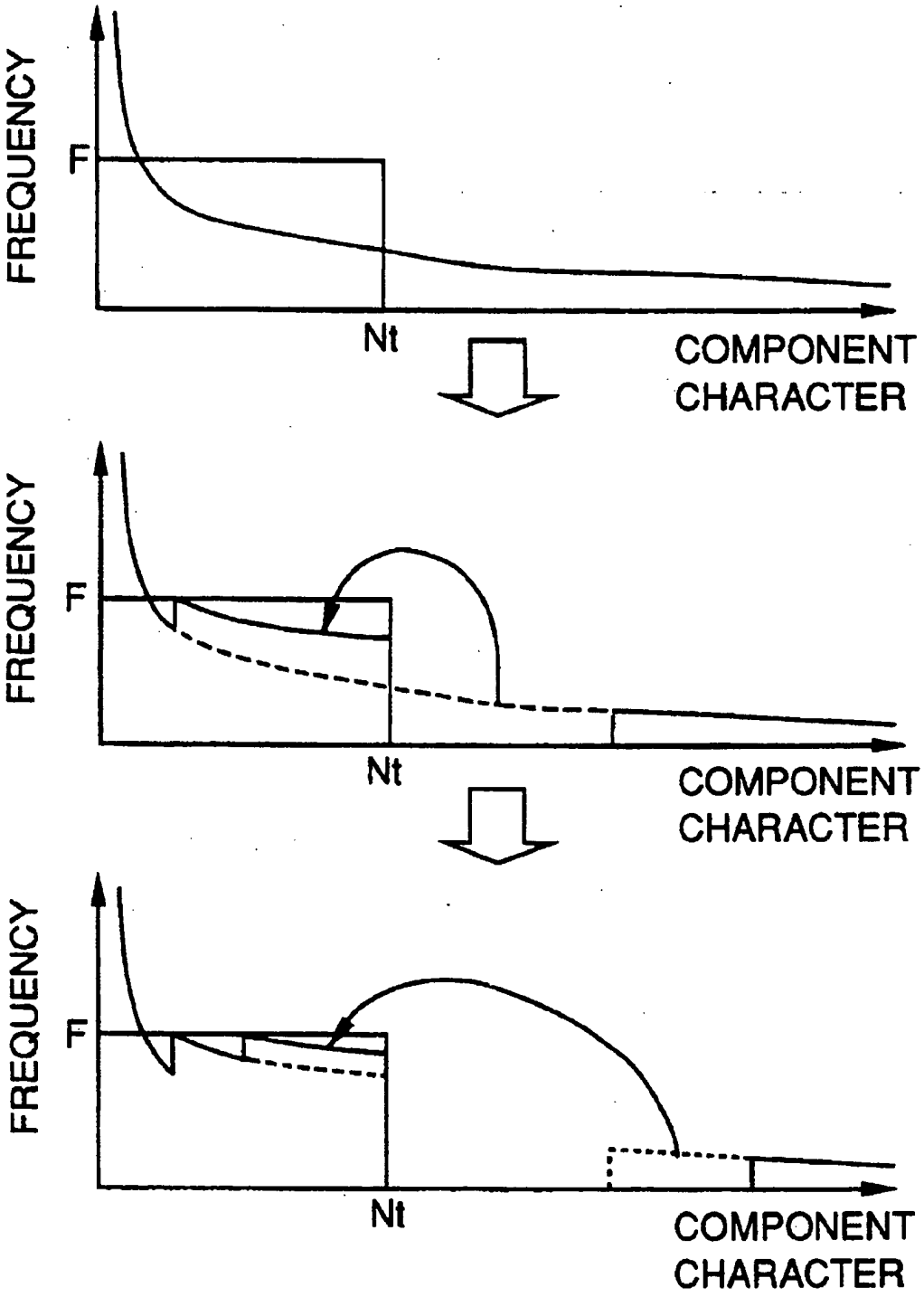

FREQUENCY
F
Nt
COMPONENT CHARACTER
FREQUENCY
F
Nt
COMPONENT CHARACTER
FREQUENCY
F
Nt
COMPONENT CHARACTER

FIG. 63

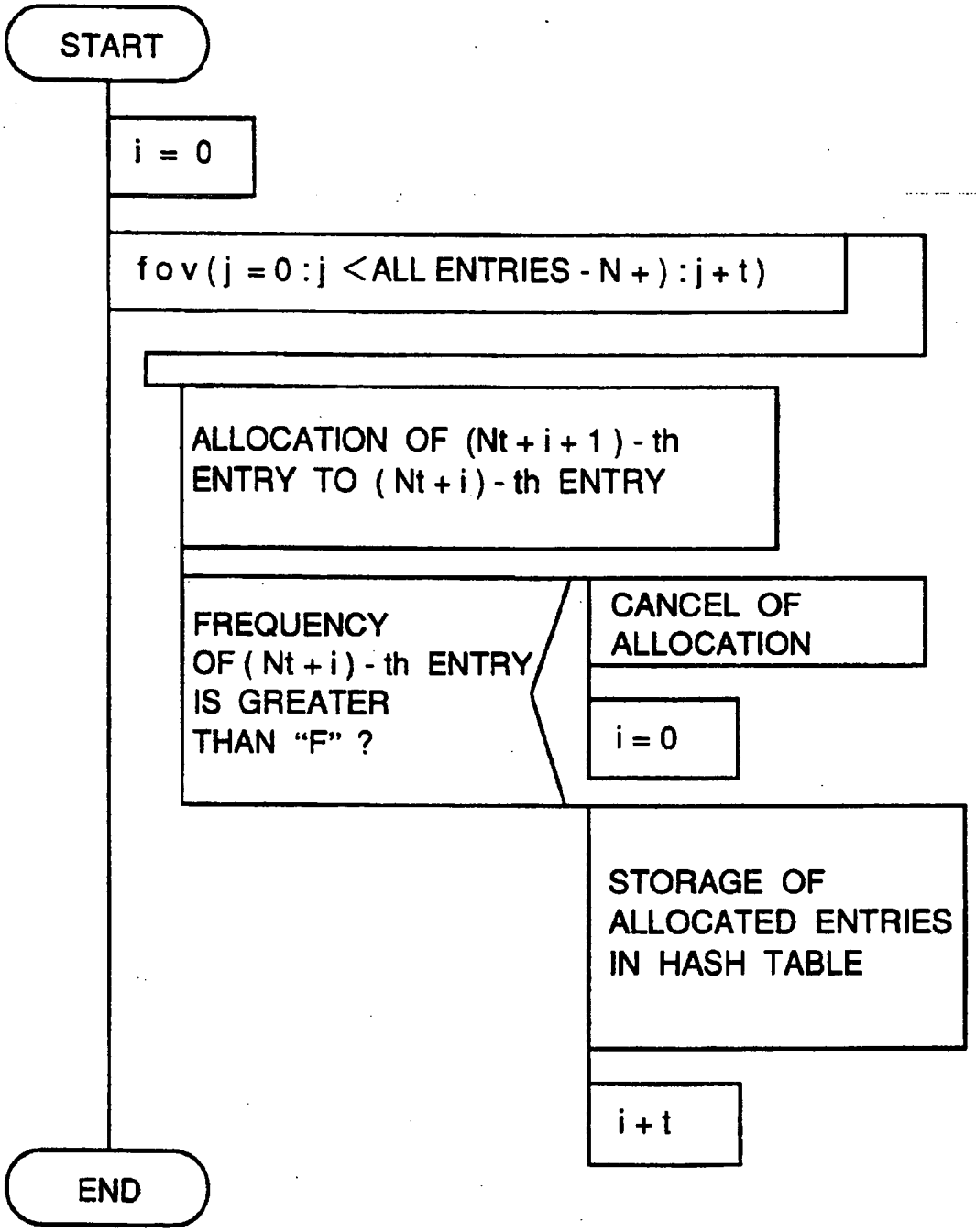

START
i = 0
f o v ( j = 0 : j < ALL ENTRIES - N + ) : j + t )
ALLOCATION OF (Nt + i + 1 ) - th ENTRY TO ( Nt + i ) - th ENTRY
FREQUENCY OF ( Nt + i ) - th ENTRY IS GREATER THAN "F" ?
CANCEL OF ALLOCATION
i = 0
STORAGE OF ALLOCATED ENTRIES IN HASH TABLE
i + t
END

FIG. 64

| | SCODE1SCODE2 | HASH ENTRY |
|---|---|---|
| | . | . |
| | . | . |
| | (096F0A8C)H | 4032 |
| | (096F0A8D)H | 167 |
| | (096F0A8E)H | 1680 |
| 「経済」→ | (096F0A8F)H | 34 |
| | (096F0A90)H | 2687 |
| | (096F0A91)H | 2948 |
| | (096F0A92)H | 862 |
| | . | . |
| | . | . |

FIG. 65

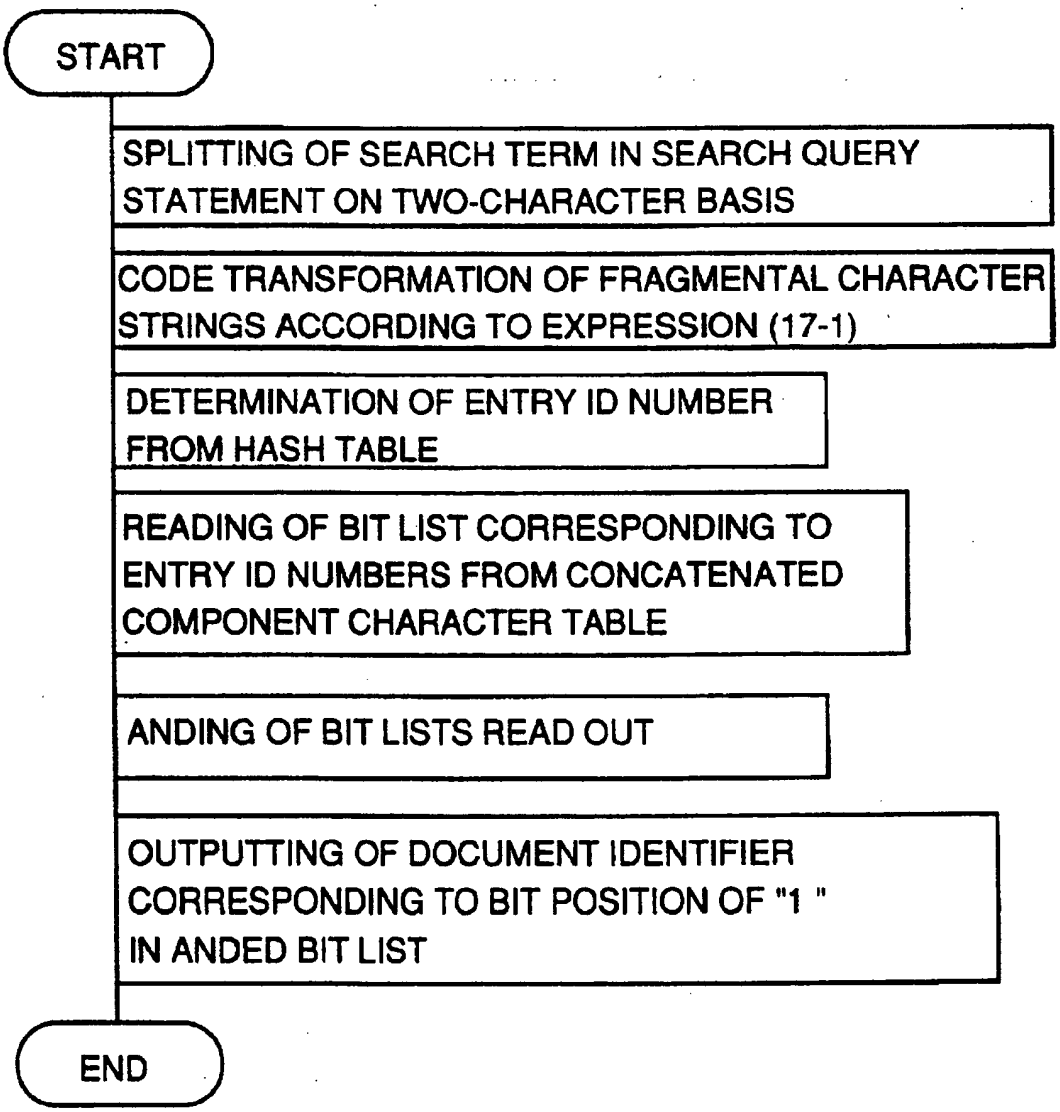

START
SPLITTING OF SEARCH TERM IN SEARCH QUERY STATEMENT ON TWO-CHARACTER BASIS
CODE TRANSFORMATION OF FRAGMENTAL CHARACTER STRINGS ACCORDING TO EXPRESSION (17-1)
DETERMINATION OF ENTRY ID NUMBER FROM HASH TABLE
READING OF BIT LIST CORRESPONDING TO ENTRY ID NUMBERS FROM CONCATENATED COMPONENT CHARACTER TABLE
ANDING OF BIT LISTS READ OUT
OUTPUTTING OF DOCUMENT IDENTIFIER CORRESPONDING TO BIT POSITION OF "1 " IN ANDED BIT LIST
END

FIG. 66

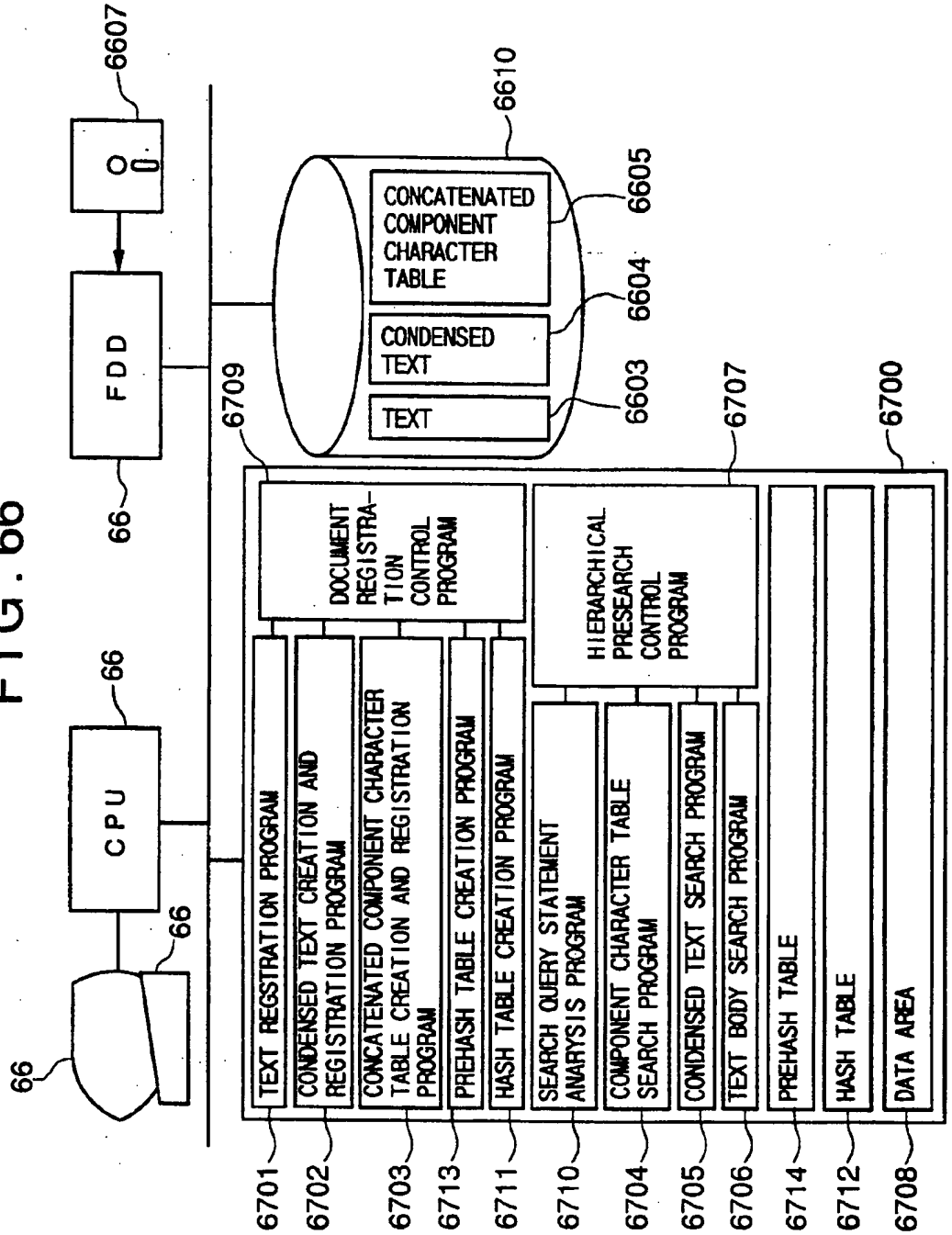

6607
O
FDD
66
CPU
66
66
66
6610
CONCATENATED COMPONENT CHARACTER TABLE
6605
CONDENSED TEXT
6604
TEXT
6603
6709
6707
6700
DOCUMENT REGISTRATION CONTROL PROGRAM
HIERARCHICAL PRESEARCH CONTROL PROGRAM
TEXT REGSTRATION PROGRAM
CONDENSED TEXT CREATION AND REGISTRATION PROGRAM
CONCATENATED COMPONENT CHARACTER TABLE CREATION AND REGISTRATION PROGRAM
PREHASH TABLE CREATION PROGRAM
HASH TABLE CREATION PROGRAM
SEARCH QUERY STATEMENT ANARYSIS PROGRAM
COMPONENT CHARACTER TABLE SEARCH PROGRAM
CONDENSED TEXT SEARCH PROGRAM
TEXT BODY SEARCH PROGRAM
PREHASH TABLE
HASH TABLE
DATA AREA
6701
6702
6703
6713
6711
6710
6704
6705
6706
6714
6712
6708

FIG. 67

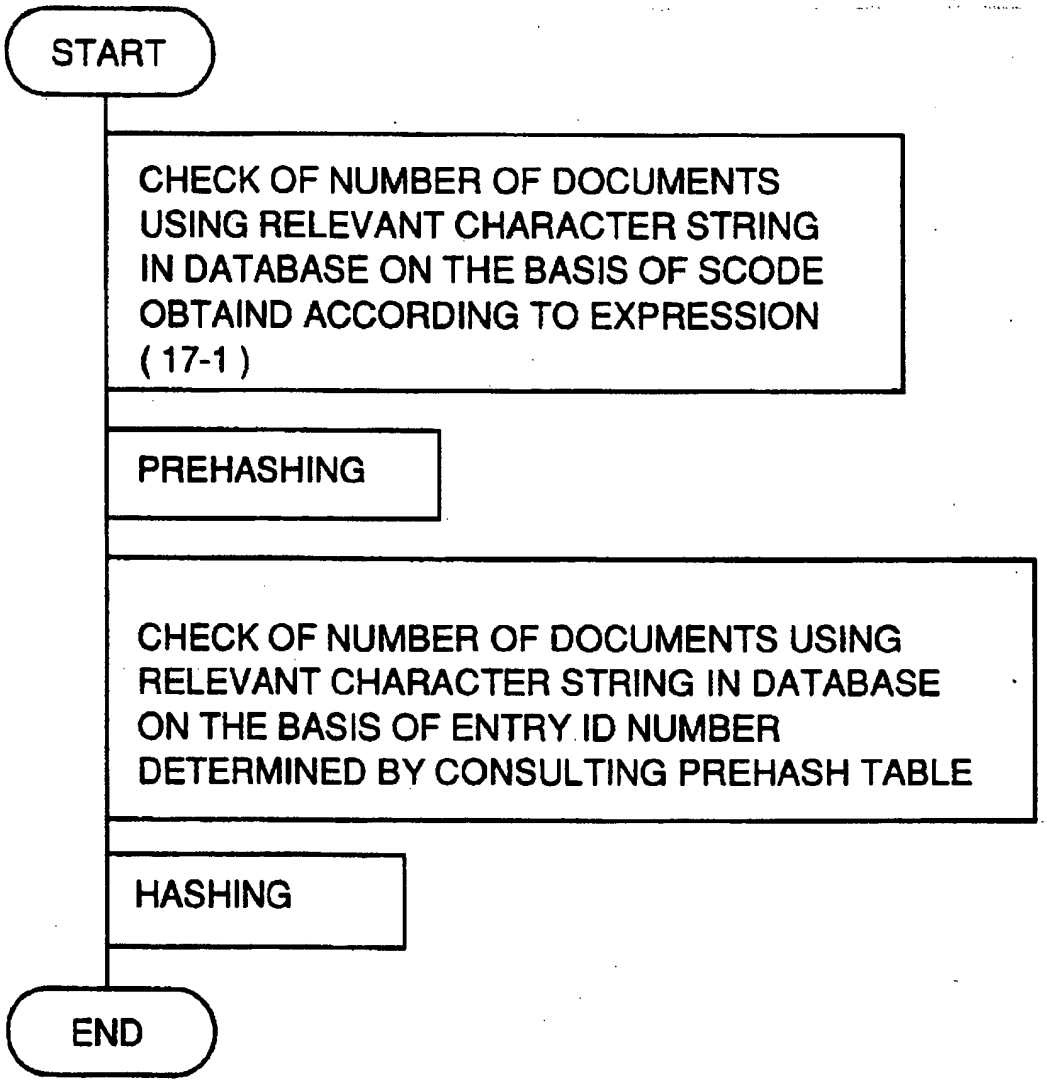
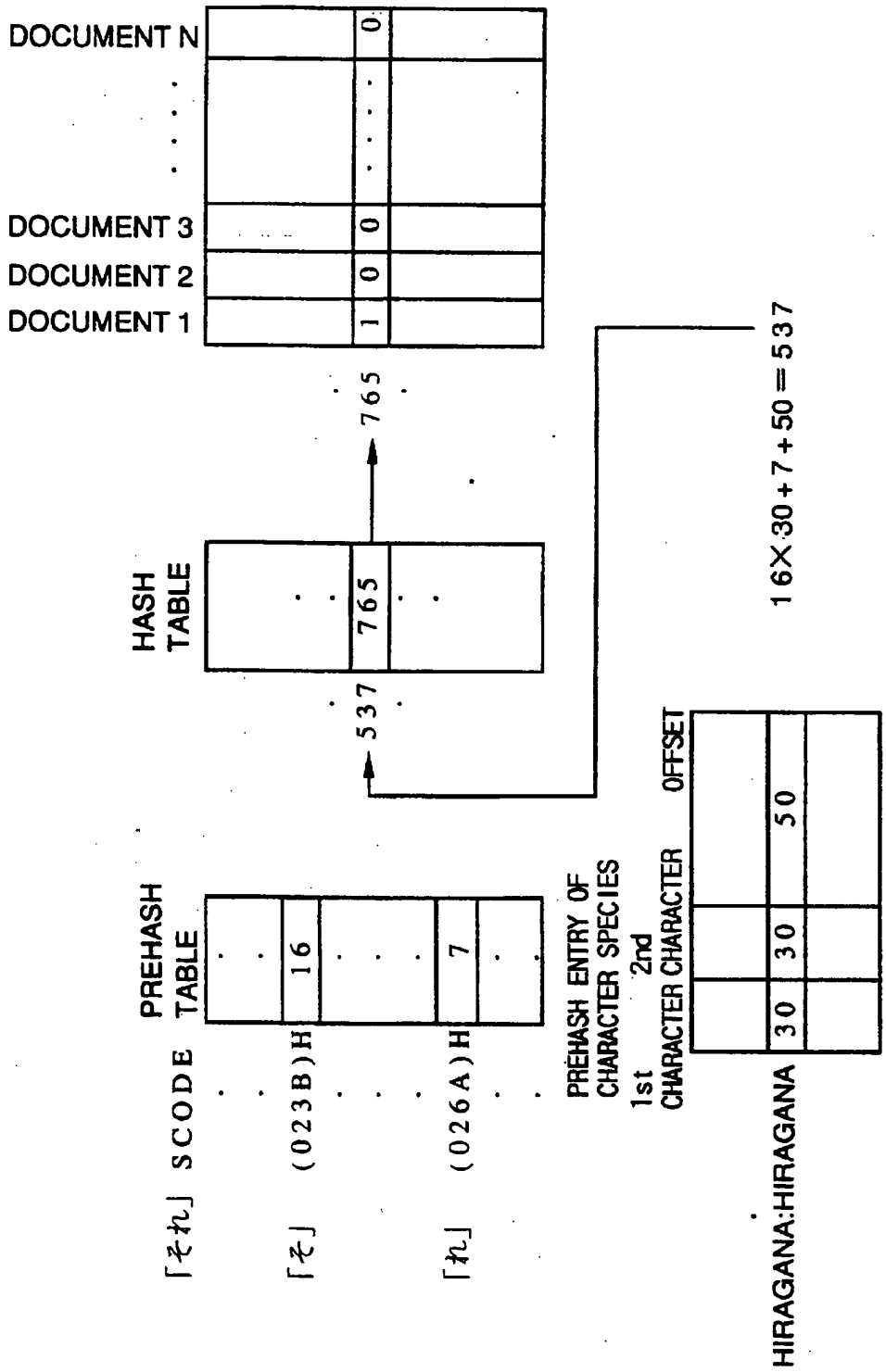

START
CHECK OF NUMBER OF DOCUMENTS USING RELEVANT CHARACTER STRING IN DATABASE ON THE BASIS OF SCODE OBTAIND ACCORDING TO EXPRESSION ( 17-1 )
PREHASHING
CHECK OF NUMBER OF DOCUMENTS USING RELEVANT CHARACTER STRING IN DATABASE ON THE BASIS OF ENTRY ID NUMBER DETERMINED BY CONSULTING PREHASH TABLE
HASHING
END

FIG. 68


DOCUMENT N
DOCUMENT 3
DOCUMENT 2
DOCUMENT 1
765
HASH TABLE
537
16×30+7+50=537
PREHASH TABLE
「それ」 SCODE
「そ」 (023B)H
「れ」 (026A)H
PREHASH ENTRY OF CHARACTER SPECIES
1st CHARACTER
2nd CHARACTER
OFFSET
HIRAGANA:HIRAGANA